Roskilde University

Basisstatistik 2

kontinuerte modeller

Larsen, Jørgen

Publication date:
1988

Document Version
Også kaldet Forlagets PDF

Citation for published version (APA):
Larsen, J. (1988). *Basisstatistik 2: kontinuerte modeller*. Roskilde Universitet. Tekster fra IMFUFA Nr. 167b

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain.
- You may freely distribute the URL identifying the publication in the public portal.

Take down policy
If you believe that this document breaches copyright please contact rucforsk@kb.dk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 05. May. 2023

TEKST NR 167b 1988

BASISSTATISTIK
2. Kontinuerte modeller

Jørgen Larsen

IMFUFA
Roskilde Universitetscenter

September 1988

TEKSTER fra

IMFUFA ROSKILDE UNIVERSITETSCENTER
INSTITUT FOR STUDIET AF MATEMATIK OG FYSIK SAMT DERES FUNKTIONER I UNDERVISNING, FORSKNING OG ANVENDELSER

Kapitel 9

Kontinuerte fordelinger; eksponentialfordelingen

Når man opbygger en statistisk model for et datamateriale, tilstræber man ofte at fundamentale træk i datamaterialets struktur genspejles i modellen. Et fundamentalt træk er *arten af observationerne*:

- I nogle situationer er elementar-observationerne resultaterne af en *klassifikation* af et bestemt antal individer i et bestemt antal klasser (er personen rød-, lys-, brun- eller sorthåret). I så fald benyttes ofte multinomialfordelingsmodeller (eller binomialfordelingsmodeller hvis der kun er to klasser).

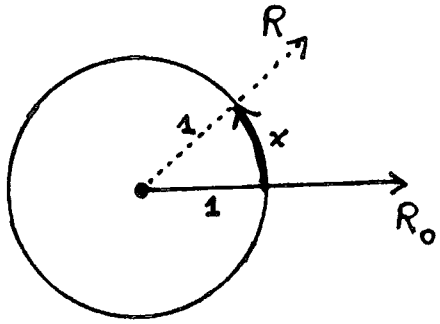
- I andre situationer er elementar-observationerne *antal* (antal lungekræfttilfælde i en treårsperiode, antal biller på en halv kvadratmeter mark etc.). Der er ikke nogen principiel øvre grænse for størrelsen af disse antal. Her kan en Poissonfordelingsmodel komme på tale.

- I atter andre situationer måler man størrelser på en *kontinuert skala*[1], f.eks. tider (ventetider), højder, længder, masser, koncentrationer osv.[2]

[1] At skalaen er *kontinuert* betyder, at man principielt kan få "enhver" værdi. Der er f.eks. ikke noget i vejen for, at en person på et givet tidspunkt kan veje 71.3478 kg, hvorimod det er principielt umuligt at personen kan have 2.8 børn!

[2] Bemærk i øvrigt, at kontinuerte observationer meget ofte er *benævnte* størrelser.

Figur 9.1: Eksempel 9.1: Bestemmelse af vinklen x fra retning R_0 til retning R: De to retninger afsættes ud fra et punkt, der desuden er centrum for en enhedscirkel. Længden af cirkelbuen *fra* R_0s skæring *til* Rs skæring med cirklen, målt mod uret, er et enkelt tal der fuldstændig specificerer retningen R.



Dette og de følgende kapitler beskæftiger sig med eksempler på statistiske modeller for kontinuerte observationer.

Eksempel 9.1. *Fugles flugt*

Vi begynder med et simpelt eksempel. Antag at man udfører et forsøg der består i at slippe en tilfangetagen fugl løs idet man observerer, i hvilken retning den flyver bort. Det er tænkeligt at fuglen har nogle foretrukne retninger at forsvinde i, men vi holder os til det simpleste og antager, at fuglen slet og ret vælger en retning tilfældigt, således at alle retninger er lige sandsynlige. Hvad vil det nærmere sige?

Vi formaliserer problemet lidt. Vi vil gerne kunne specificere flugtretningen R ved ét (eller flere) tal. Hvis vi fastlægger en "reference-retning" R_0, kan vi specificere retningen R ved at angive, hvor stor en vinkel x der er fra R_0 til R, se Figur 9.1. Størrelsen x bliver et tal mellem 0 og 2π.

I vores formalisering er det nu sådan at fuglen vælger et x tilfældigt således at alle x-værdier er lige sandsynlige. Eksempelvis er det lige så sandsynligt at få en x-værdi mellem 1.0 og 1.1 som mellem 1.1 og 1.2, og derfor må sandsynligheden for at få en værdi mellem 1.0 og 1.2 være det dobbelte af sandsynligheden for at

få en værdi mellem 1.0 *og* 1.1. *Ved at udbygge dette ræsonnement lidt finder man, at der må gælde at sandsynligheden for at få en* x*-værdi mellem* a *og* b *må være proportional med længden af intervallet fra* a *til* b*, mere præcist må sandsynligheden være*

$$\frac{b-a}{2\pi}$$

når $0 \leq a \leq b \leq 2\pi$. *Hvis vi introducerer en stokastisk variabel* X *der skal stå for "den retning som fuglen nu tilfældigvis vælger", så har vi altså denne simple sandsynlighedsmodel for* X*:*

$$\mathrm{P}(a < X \leq b) \;=\; \frac{b-a}{2\pi}$$

når $0 \leq a \leq b \leq 2\pi$.

Man kan godt spørge om sandsynligheden for at X *antager en bestemt værdi* x_0 *- og få et svar: Da udsagnet* $X = x_0$ *medfører udsagnet* $x_0 - h < X \leq x_0 + h$*, ligegyldigt hvilket positivt tal* h *vi vælger, så er*

$$\begin{aligned} \mathrm{P}(X = x_0) &\leq \mathrm{P}(x_0 - h < X \leq x_0 + h) \\ &= \frac{(x_0 + h) - (x_0 - h)}{2\pi} \\ &= h/\pi \; . \end{aligned}$$

Sandsynligheden for at $X = x_0$ *er således mindre end* h/π *for ethvert nok så lille tal* h*, dvs.* $\mathrm{P}(X = x_0) \leq 0$*, og da sandsynligheder på den anden side aldrig er negative, så er svaret altså, at* $\mathrm{P}(X = x_0)$ *er 0.*

X *er et eksempel på en kontinuert stokastisk variabel. Fordelingen af en kontinuert stokastisk variabel kan man* ikke *specificere ved hjælp af en sandsynlighedsfunktion der angiver sandsynligheden for hvert enkelt muligt udfald, fordi dén sandsynlighed er altid 0. I stedet specificerer man fordelingen ved hjælp af den såkaldte* tæthedsfunktion *der angiver "hvor tæt sandsynlighedsmassen ligger forskellige steder på talaksen", målt som* sandsynlighed pr. intervallængde. *Ovenfor fandt vi sandsynligheden for at* X *ligger i et interval af længde* $2h$ *omkring* x_0 *til at være* h/π*, hvilket betyder at sandsynlighedstætheden i punktet* x_0 *er lig* $1/2\pi$.

Tæthedsfunktionen for X er dermed

$$f(x) = \begin{cases} 1/2\pi & \text{når } 0 \leq x \leq 2\pi \\ 0 & \text{ellers .} \end{cases}$$

Den sandsynlighedsfordeling som X følger hedder ligefordelingen *på intervallet fra 0 til 2π.* □

Læseren henvises til *Grundbegreber i Sandsynlighedsregningen* for en oversigt over det beskrivelsesapparat man benytter sig af i forbindelse med kontinuerte sandsynlighedsfordelinger.

Eksponentialfordelingen

Som et noget større eksempel vil vi vise hvordan *eksponentialfordelingen* naturligt kan komme på tale, og vi vil give et meget simpelt eksempel på statistisk analyse af eksponentialfordelte observationer.

En ventetidsfordeling

I Kapitel 7 udledte vi en model for antallet af begivenheder i et bestemt tidsinterval af længde t, under antagelse af at begivenhederne indtræffer tilfældigt og uafhængigt af hverandre. Resultatet var (jf. Resumé 6), at antallet af begivenheder måtte være Poissonfordelt med parameter λt, hvor λ er den intensitet hvormed begivenhederne indtræffer. Vi vil nu udlede en anden konsekvens af de samme antagelser, vi vil nemlig bestemme fordelingen af ventetiden til den næste begivenhed.

Først bestemmer vi fordelingen af ventetiden T_1 fra et fast valgt tidspunkt t_0 til den førstkommende begivenhed. Vi vil bestemme sandsynligheden for at der går mere end tiden t før der indtræffer en begivenhed, dvs. sandsynligheden for at $T_1 > t$.

At $T_1 > t$ er det samme som at der forekommer 0 begivenheder i tidsintervallet fra t_0 til t_0+t. Da vi véd at antal begivenheder i intervallet $]t_0, t_0+t]$ er Poissonfordelt med parameter λt, så er sandsynligheden for at der sker 0 begivenheder

$$\frac{(\lambda t)^0}{0!}\exp(-\lambda t) \;=\; \exp(-\lambda t)\,.$$

Vi har dermed fundet at

$$\mathrm{P}(T_1 > t) \;=\; \exp(-\lambda t)\,.$$

Fordelingsfunktionen for T_1 er da

$$\begin{aligned} F(t) &= \mathrm{P}(T_1 \le t) \\ &= 1 - \mathrm{P}(T_1 > t) \\ &= 1 - \exp(-\lambda t), \quad t \ge 0\,, \end{aligned}$$

og *tæthedsfunktionen* for T_1 er

$$\begin{aligned} f(t) &= F'(t) \\ &= \lambda\exp(-\lambda t), \quad t \ge 0\,. \end{aligned} \tag{9.1}$$

Den sandsynlighedsfordeling på den positive halvakse der har (9.1) som tæthed hedder *eksponentialfordelingen med parameter* λ.

Vi er således nået frem til, at *ventetiden fra* t_0 *til den næste begivenhed er eksponentialfordelt med parameter* λ. Eftersom begivenhederne indtræffer uafhængigt af hverandre, er dette resultat også rigtigt i den situation hvor man ved, at der tilfældigvis indtraf en begivenhed netop til tid t_0. Derfor gælder også, at *ventetiden mellem en begivenhed og den næste er eksponentialfordelt med parameter* λ.

Hvad der sker af begivenheder *efter* t_0 er ganske uafhængigt af, hvad der er sket *før* t_0, så der gælder, at selv om det er kendt hvornår begivenheden før den begivenhed til tid t_0 indtraf, så ændrer det ikke noget på fordelingen af tidsrummet mellem begivenheden til tid t_0 og den næste begivenhed. Det medfører, at de to ventetider mellem de tre på hinanden følgende begivenheder er *stokastisk uafhængige* identisk eksponentialfordelte variable. Ved at bygge videre på ræsonnementet fås, at ventetiderne mellem vilkårligt mange på hinanden følgende begivenheder er *uafhængige* identisk eksponentialfordelte størrelser.

Egenskaber ved eksponentialfordelingen

Eksponentialfordelingen med parameter λ har *fordelingsfunktion*

$$F(t) = \begin{cases} 1 - \exp(-\lambda t) & \text{for } t \geq 0 \\ 0 & \text{for } t < 0 \, , \end{cases}$$

og *tæthedsfunktion*

$$f(t) = \begin{cases} \lambda \exp(-\lambda t) & \text{for } t \geq 0 \\ 0 & \text{for } t < 0 \, . \end{cases}$$

Middelværdien af en stokastisk variabel T som er eksponentialfordelt med parameter λ er

$$\begin{aligned} \mathrm{E}\, T &= \int_0^{+\infty} t \, \lambda \exp(-\lambda t) \, dt \\ &= 1/\lambda \, , \end{aligned} \tag{9.2}$$

og *variansen* er

$$\begin{aligned} \mathrm{Var}\, T &= \mathrm{E}(T - \mathrm{E}\, T)^2 \\ &= \mathrm{E}\, T^2 - (\mathrm{E}\, T)^2 \\ &= \int_0^{+\infty} t^2 \, \lambda \exp(-\lambda t) \, dt \; - \; 1/\lambda^2 \, ; \end{aligned}$$

ved partiel integration finder man at integralet er lig med $2/\lambda^2$, så

$$\mathrm{Var}\, T = 1/\lambda^2 \, .$$

Poissonprocessen

Når begivenhederne indtræffer som beskrevet i Resumé 6 siger man, at de indtræffer som efter en *Poissonproces* med intensitet λ. Da gælder at

- antallene af begivenheder i ikke-overlappende tidsintervaller er uafhængige og Poissonfordelte, således at Poissonfordelingen hørende til et interval af længde t har parameter λt,

- ventetider mellem på hinanden følgende begivenheder er uafhængige eksponentialfordelte størrelser med parameter λ, dvs. med middelværdi $1/\lambda$,
- ventetiden fra et fast tidspunkt til den førstkommende begivenhed er ligeledes eksponentialfordelt med parameter λ.

Man taler om en Poisson*proces* for at henlede opmærksomheden på, at der er tale om noget der forløber i tiden. Her er nogle eksempler på situationer fra virkeligheden, hvor en Poissonproces kan komme på tale som model: kunders ankomst til en supermarkedskasse; telefonopkald til en given telefoncentral; biltrafik på en landevej.

Eksempel 9.2. ***Biler på en landevej***

Når man vil måle hvor trafikeret en vej er, kan man gøre det at man registrerer det nøjagtige tidspunkt for hver enkelt bils passage af en bestemt stribe tværs over vejen. I Tabel 9.1 er vist nogle målinger fra en sådan trafiktælling et sted langs hovedvej A1. Måleperiodens begyndelsestidspunkt er $t = 0$; tabellen viser ankomsttidspunkterne for alle de biler der passerede i løbet af det første kvarter.

Vi søger en sandsynlighedsmodel der kan beskrive trafikken. Hvis vejstrækningen er langt fra lyskurve, byer og færgehavne, kan man måske tænke sig, at bilerne kommer nogenlunde tilfældigt, måske ligefrem at de kommer som efter en Poissonproces.

Hvis *de kommer som en Poissonproces, så er ventetiderne mellem på hinanden følgende biler eksponentialfordelte. Vi kan derfor få en kontrol af Poissonproces-antagelsen ved at se efter om det er rimeligt at antage, at ventetiderne er eksponentialfordelte. Derfor udregnes ventetiderne mellem de på hinanden følgende biler, se Tabel 9.2.* □

Der er nu behov for metoder til

1. at estimere den ukendte parameter λ i eksponentialfordelingen, under forudsætning af at data faktisk kan beskrives ved en eksponentialfordeling.

Tabel 9.1: Biler på en landevej: Ankomsttidspunkter (i sekunder) for de biler der passerede tælleapparatet i tidsrummet fra $t = 0$ til $t = 900$ sek.

25.0	248.2	518.8	677.4	790.8
68.7	280.8	520.2	678.7	794.6
70.2	285.2	526.7	685.4	798.5
110.3	292.8	530.0	690.5	818.0
111.2	338.7	535.6	693.1	819.3
113.0	343.1	565.5	723.9	830.0
128.1	377.5	577.2	726.4	843.6
182.4	385.1	585.4	732.6	845.6
198.5	454.8	597.6	746.4	848.4
203.6	467.1	614.4	767.9	849.4
205.3	498.1	622.4	773.9	851.2
207.0	507.4	626.2	782.9	862.0
242.8	511.9	669.2	787.8	894.4

Tabel 9.2: Biler på en landevej: Ventetider (i sekunder) mellem 64 på hinanden følgende biler.

43.7	32.6	1.4	1.3	3.8
1.5	4.4	6.5	6.7	3.9
40.1	7.6	3.3	5.1	19.5
0.9	45.9	5.6	2.6	1.3
1.8	4.4	29.9	30.8	10.7
15.1	34.4	11.7	2.5	13.6
54.3	7.6	8.2	6.2	2.0
16.1	69.7	12.2	13.8	2.8
5.1	12.3	16.8	21.5	1.0
1.7	31.0	8.0	6.0	1.8
1.7	9.3	3.8	9.0	10.8
35.8	4.5	43.0	4.9	32.4
5.4	6.9	8.2	3.0	

2. at vurdere om den under 1 fittede, bedste eksponentialfordeling nu også er god nok til at beskrive den variation der er i talmaterialet.

Modelfunktion og likelihoodfunktion i kontinuerte modeller

I forbindelse med diskrete modeller opererer vi med en såkaldt *modelfunktion* der blot er sandsynlighedsfunktionen betragtet som en funktion af både observationer og parametre, jf. f.eks. Resumé 2. På ganske tilsvarende måde er modelfunktionen for en kontinuert statistisk model simpelthen sandsynlighedstæthedsfunktionen betragtet som en funktion af dels observationerne, dels de ukendte parametre.

Likelihoodfunktionen fås som altid ud fra modelfunktionen ved at indsætte de faktiske observationer på observationsvariablenes plads og betragte udtrykket som en funktion af parametrene alene.

I eksemplet er tæthedsfunktionen for en enkelt ventetid Y eksponentialfordelingstætheden $\lambda \exp(-\lambda y)$, $y \geq 0$. Den simultane tæthedsfunktion for n $(= 64)$ uafhængige identisk eksponentialfordelte størrelser er

$$\prod_{i=1}^{n} \lambda \exp(-\lambda y_i) \;=\; \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} y_i\right)$$

når $y_i \geq 0$ for alle i. Modelfunktionen er altså

$$\begin{aligned} f(y_1, y_2, \ldots, y_n; \lambda) &= \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} y_i\right) \\ &= \lambda^n \exp(-\lambda v)\,, \end{aligned}$$

hvor

$$v \;=\; \sum_{i=1}^{n} y_i$$

er den totale observerede ventetid[3].

[3]Bemærk i øvrigt at v simpelt hen er differensen mellem den sidste og den første tidsregistrering, i taleksemplet altså $v = 894.4 - 25 = 869.4$ sek.

Likelihoodfunktionen bliver således

$$L(\lambda) = \lambda^n \exp(-\lambda v) ,$$

og log-likelihoodfunktionen er

$$\ln L(\lambda) = n \ln \lambda - \lambda v .$$

For at bestemme dennes maksimumspunkt undersøges hvornår $(\ln L)'$ er lig 0:

$$(\ln L)'(\lambda) = \frac{n}{\lambda} - v$$

er 0 netop når λ er lig n/v. Da $(\ln L)''$ er negativ, er n/v et maksimumspunkt. Følgelig er maksimaliseringsestimatet for λ

$$\widehat{\lambda} = \frac{n}{v} ,$$

nemlig antallet af ankomster divideret med den totale observerede ventetid. I bileksemplet er

$$\begin{aligned} \widehat{\lambda} &= \frac{64}{869.4 \text{ sek}} \\ &= 0.0736 \text{ sek}^{-1} ; \end{aligned}$$

bilerne passerer altså med en intensitet på ca. 0.07 biler pr. sekund, svarende til en middelventetid mellem to på hinanden følgende biler på ca. $1/\widehat{\lambda} = 14$ sekunder, jf. (9.2) side 162.

Hermed har vi fået løst problem 1 på side 163.

Histogrammer, empiriske fordelingsfunktioner, fraktildiagrammer

Vi har set hvordan man i princippet kan bestemme den bedst mulige fordeling af en given type. Spørgsmålet er så, om den er god nok!

Der findes adskillige mere eller mindre sofistikerede metoder som man kan tage i anvendelse for at vurdere hvor godt data beskrives af en given type fordeling. De forskellige metoder har deres berettigelse derved, at de kan være gode til at afsløre forskellige slags afvigelser

fra den formodede model. Hvis man ikke på forhånd netop er på jagt efter helt bestemte typer af afvigelser, er det ganske afgjort en fordel at begynde med, og måske holde sig til, simple grafiske metoder der "blot" viser, hvordan tallene fordeler sig. Sådanne grafiske metoder kan gå ud på at

- tegne histogrammer og sammenligne dem med den fittede teoretiske tæthedsfunktion,
- tegne den empiriske fordelingsfunktion og sammenligne den med den fittede teoretiske fordelingsfunktion,
- tegne et fraktildiagram ('probability plot').

Histogrammer

Et histogram er en art empirisk tæthedsfunktion. Man fremstiller et histogram på den måde, at man inddeler abscisseaksen i nogle passende intervaller[4], hvorefter man for hvert interval tegner et rektangel hvis "grundflade" er intervallet og hvis areal er lig med den observerede brøkdel af observationer i det pågældende interval[5].

Histogrammet skal ligne den fittede teoretiske fordelings tæthedsfunktion, så den kan man passende tegne ind på den samme figur. Figur 9.2 er et eksempel på et histogram over bil-tallene. Når man skal udarbejde et histogram (eller en af de øvrige grafiske fremstillinger i dette afsnit) ved "håndkraft", er det gerne en lettelse at begynde med at danne de *ordnede observationer*, hvilket blot er observationerne ordnet i (voksende) rækkefølge; hvis observationerne hedder $y_1, y_2, \ldots, y_n$, plejer man at betegne de ordnede observationer

$$y_{(1)}, y_{(2)}, \cdots, y_{(n)} \cdot$$

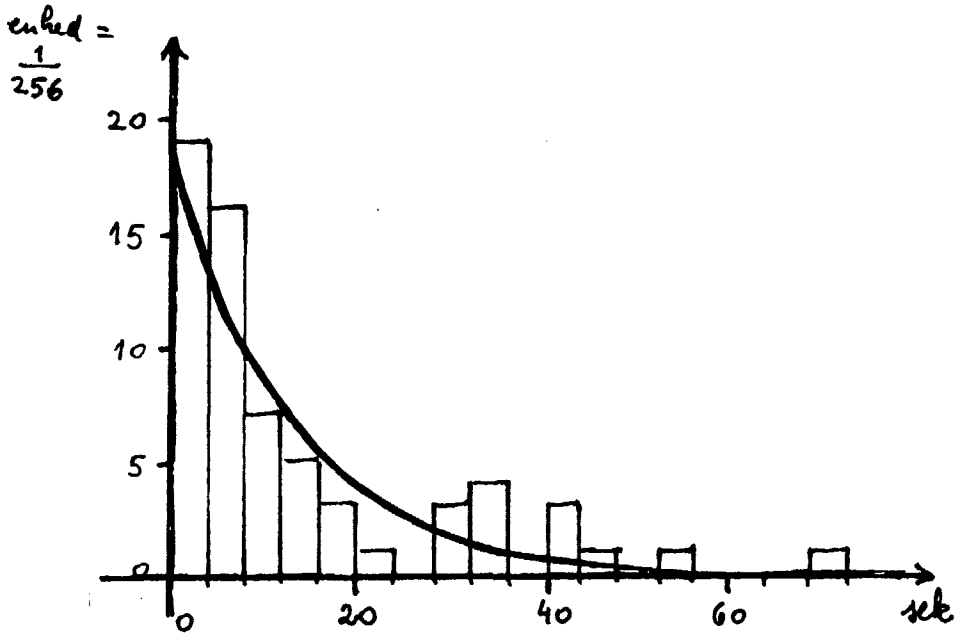
De ordnede bil-observationer er vist i Tabel 9.3.

Ved udarbejdelsen af histogrammet kan det være lidt af et kunststykke at vælge den rigtige intervalinddeling, som bevirker at fluktuationerne bliver passende udjævnet uden at tæthedens form bliver alt for

[4]Intervallerne behøver ikke nødvendigvis at have samme længde, men det letter normalt figurens forståelighed hvis de *har* samme længde.

[5]For at undgå problemer med eventuelle observationer i interval-endepunkterne kan man "snyde" og vælge intervallerne således at ingen af observationerne ligger i et intervalendepunkt

Figur 9.2: Biler på en landevej: Histogram (svarende til klassebredden 4 sekunder) over ventetiderne mellem på hinanden følgende biler, samt tætheden for eksponentialfordelingen med parameter $\widehat{\lambda} = 0.736\,\text{sek}^{-1}$.



Tabel 9.3: Biler på en landevej: de ordnede ventetids-observationer.

0.9	2.8	5.6	10.7	30.8
1.0	3.0	6.0	10.8	31.0
1.3	3.3	6.2	11.7	32.4
1.3	3.8	6.5	12.2	32.6
1.4	3.8	6.7	12.3	34.4
1.5	3.9	6.9	13.6	35.8
1.7	4.4	7.6	13.8	40.1
1.7	4.4	7.6	15.1	43.0
1.8	4.5	8.0	16.1	43.7
1.8	4.9	8.2	16.8	45.9
2.0	5.1	8.2	19.5	54.3
2.5	5.1	9.0	21.5	69.7
2.6	5.4	9.3	29.9	

udjævnet. Hvis intervallerne er for korte bliver fluktuationerne ikke udglattet nok, er de for lange sker der en for stor udjævning af tæthedens form.

Man kan naturligvis godt opskrive definitionen på et histogram lidt mere formelt: Opgaven består i at udarbejde et histogram for observationssættet $y_1, y_2, \ldots, y_n$. Den løses således:

1. I det område hvor observationerne falder vælges delepunkter (der som regel bør være ækvidistante) $x_0 < x_1 < x_2 < \ldots < x_m$, hvor x_0 er mindre end den mindste og x_m større end den største af y-observationerne.

2. Bestem antallet n_j af y-er i det j-te interval som er $]x_{j-1}, x_j]$.

3. Definer den stykkevis konstante funktion

$$h(y) = \begin{cases} \dfrac{n_j/n}{x_j - x_{j-1}} & \text{når } y \in]x_{j-1}, x_j] \\ 0 & \text{når } y \leq x_0 \text{ eller } y > x_m \end{cases}$$

Så er histogrammet (svarende til den valgte inddeling) over observationerne $y_1, y_2, \ldots, y_n$ grafen for h.

Empirisk fordelingsfunktion, fraktildiagram

Et histogram er en slags empirisk *tæthed*sfunktion, som man kan sammenligne med den fittede teoretiske tæthedsfunktion. Man kan imidlertid også tegne en empirisk *fordelings*funktion for at søge at sammenligne den med den fittede teoretiske fordelingsfunktion.

Fordelingsfunktionen F for en stokastisk variabel Y er jo

$$F : y \mapsto \mathrm{P}(Y \leq y) ,$$

så den empiriske fordelingsfunktion må skulle være en funktion der til hvert y angiver hvor stor en brøkdel af observationerne der befinder sig til venstre for y. Når der foreligger observationer $y_1, y_2, \ldots, y_n$ defineres den empiriske fordelingfunktion $\widehat{F}$ ved

$$\widehat{F}(y) = \frac{1}{n}\Big((\text{antal obs.} < y) + \tfrac{1}{2} \times (\text{antal obs.} = y)\Big) . \quad (9.3)$$

$\widehat{F}$ bliver en stykkevis konstant funktion der har spring i punkterne $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$ (dvs. i punkterne $y_1, y_2, \ldots, y_n$); funktionsværdien i et springpunkt er middeltallet af grænseværdierne fra venstre og fra højre. En empirisk fordelingsfunktion ser derfor i princippet ud som vist på Figur 9.3[6].

For at vurdere hvor godt den empiriske fordelingsfunktion $\widehat{F}$ og den fittede teoretiske fordelingsfunktion F ligner hinanden, bør man indtegne dem på samme figur. Almindeligvis vil man kun afsætte $\widehat{F}$-værdier svarende til $y = y_{(i)}$, $i = 1, 2, \ldots, n$. Man skal således vurdere, hvordan punkterne $(y_{(i)}, \widehat{F}(y_{(i)}))$ ligger i forhold til grafen for F – ideelt skal de fordele sig tilfældigt på en eller anden måde omkring denne graf.

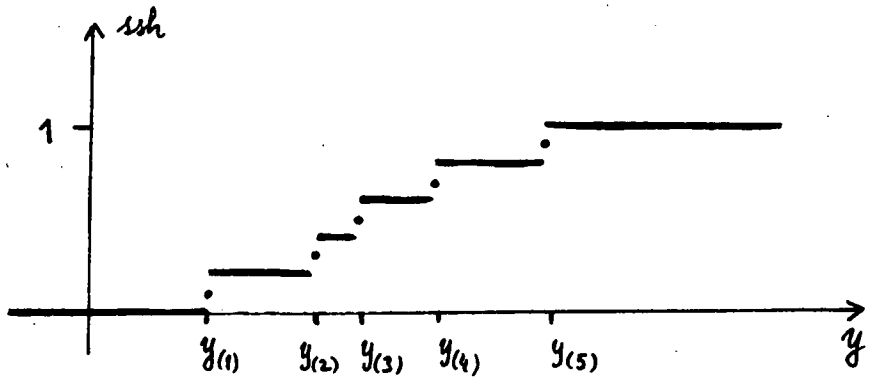
Nu er det imidlertid ikke så let at vurdere, hvordan punkter fordeler sig omkring en krum kurve, men ved et snedigt trick kan man omforme problemet til et problem hvor opgaven er at vurdere, om nogle punkter fordeler sig omkring en *ret linie*. Her viser vi, hvori tricket består når det drejer sig om eksponentialfordelingen.

[6] Ud fra histogrammet h kan man danne en anden udgave af den empiriske fordelingsfunktion, nemlig

$$\widehat{F}(y) = \int_{-\infty}^{y} h(x)\,dx .$$

Denne funktion er kontinuert og stykkevis lineær.

Figur 9.3: En empirisk fordelingsfunktion ser i princippet således ud. Den er stykkevis konstant, og dens værdier i et springpunkt er middeltallet af grænseværdierne fra venstre og fra højre.



I ventetidseksemplet formodes det at punkterne $(y_{(i)}, \widehat{F}(y_{(i)}))$ fordeler sig omkring grafen for fordelingsfunktionen for eksponentialfordelingen med en parameter λ der estimeres til $\widehat{\lambda} = 0.0736$ sek^{-1}. Denne fordelingsfunktion er $F(y) = 1 - \exp(-\widehat{\lambda}y)$ der også kan skrives som $F_0(\widehat{\lambda}y)$, hvor

$$F_0(y) = 1 - \exp(-y)$$

er fordelingsfunktionen for eksponentialfordelingen med $\lambda = 1$.

Den omvendte funktion til F_0 er

$$F_0^{-1} : p \mapsto -\ln(1-p) .$$

Det snedige trick er nu at transformere hele problemet med transformationen F_0^{-1}, det vil sige: I stedet for at betragte punkterne $(y_{(i)}, \widehat{F}(y_{(i)}))$ i forhold til grafen for $y \mapsto F(y) = 1 - \exp(-\widehat{\lambda}y)$, betragter vi punkterne $(y_{(i)}, F_0^{-1}(\widehat{F}(y_{(i)})))$ i forhold til grafen for $y \mapsto F_0^{-1}(F(y))$. Grunden til at det er smart er at

$$\begin{aligned} F_0^{-1}(F(y)) &= -\ln(1 - F(y)) \\ &= -\ln\left(1 - (1 - \exp(-\widehat{\lambda}y))\right) \\ &= \widehat{\lambda}y . \end{aligned}$$

Tabel 9.4: Ventetider: Den empiriske fordelingsfunktion $\widehat{F}$.

$y_{(i)}$	$\widehat{F}(y_{(i)})$	$y_{(i)}$	$\widehat{F}(y_{(i)})$	$y_{(i)}$	$\widehat{F}(y_{(i)})$	$y_{(i)}$	$\widehat{F}(y_{(i)})$
0.9	0.008	3.9	0.289	8.2	0.563	21.5	0.789
1.0	0.023	4.4	0.313	9.0	0.586	29.9	0.805
1.3	0.047	4.5	0.336	9.3	0.602	30.8	0.820
1.4	0.070	4.9	0.352	10.7	0.617	31.0	0.836
1.5	0.086	5.1	0.375	10.8	0.633	32.4	0.852
1.7	0.109	5.4	0.398	11.7	0.648	32.6	0.867
1.8	0.141	5.6	0.414	12.2	0.664	34.4	0.883
2.0	0.164	6.0	0.430	12.3	0.680	35.8	0.898
2.5	0.180	6.2	0.445	13.6	0.695	40.1	0.914
2.6	0.195	6.5	0.461	13.8	0.711	43.0	0.930
2.8	0.211	6.7	0.477	15.1	0.727	43.7	0.945
3.0	0.227	6.9	0.492	16.1	0.742	45.9	0.961
3.3	0.242	7.6	0.516	16.8	0.758	54.3	0.977
3.8	0.266	8.0	0.539	19.5	0.773	69.7	0.992

Det transformerede problem består derfor i at vurdere, om punkterne $(y_{(i)}, F_0^{-1}(\widehat{F}(y_{(i)})))$ fordeler sig omkring en ret linie gennem $(0,0)$ med hældning $\widehat{\lambda} = 0.0736$ sek^{-1}.

En tegning hvor punkterne $(y_{(i)}, F_0^{-1}(\widehat{F}(y_{(i)})))$ afsættes kaldes et *fraktildiagram*[7] (på engelsk: 'probability plot').

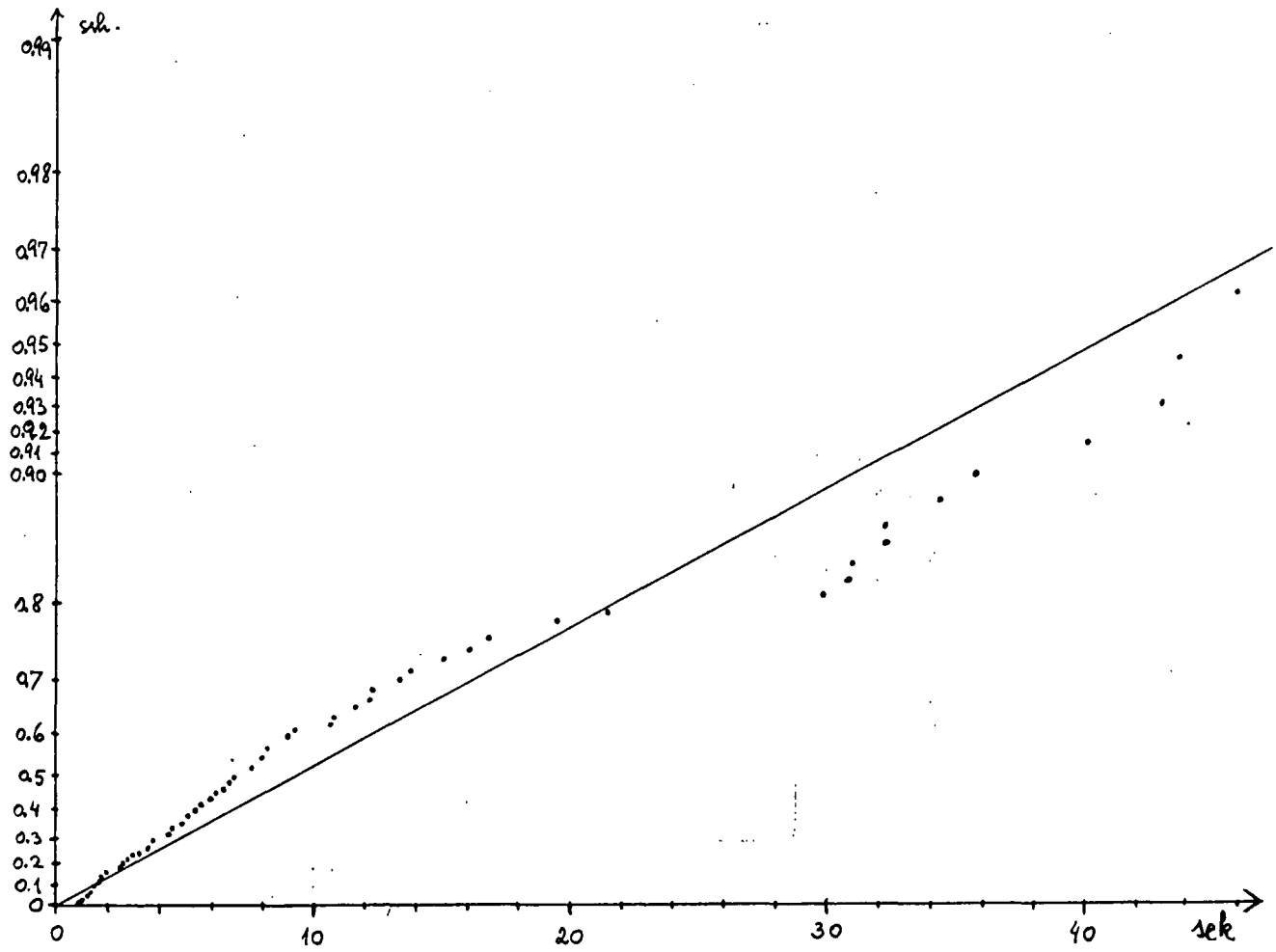
For at tegne et fraktildiagram for ventetidstallene bestemmes først den empiriske fordelingsfunktion $\widehat{F}$, se Tabel 9.4.

Man kan herefter udregne de tilsvarende værdier af $F_0^{-1}(\widehat{F}(y_{(i)})) = 1 - \ln(1 - \widehat{F}(y_{(i)})$ og afsætte punkterne $(y_{(i)}, 1 - \ln(1 - \widehat{F}(y_{(i)})))$ i et koordinatsystem, hvorved fraktildiagrammet fremkommer. Man kan dog slippe for nogle udregninger hvis man i stedet anvender *logaritmepapir*[8]. Figur 9.4 er et fraktildidagram for bil-eksemplets ventetidsfordeling.

[7] fordi tallet $F_0^{-1}(p)$ er en p-fraktil for fordelingen F_0.

[8] Benyttes papir med almindelig abscisse og logaritmisk ordinat med f.eks. to dekader, skal man vende papiret på hovedet og lade det påtrykte oprindelige "10" svare til "0", det oprindelige "9" til "0.1", det oprindelige "8" til "0.2", osv. I det derved fremkommende koordinatsystem afsættes punkterne $(y_{(i)}, \widehat{F}(y_{(i)}))$. Den teoretiske linie tegnes lettest som linien gennem $(0,0)$ og $(y, 1 - \exp(-\widehat{\lambda}y))$ for et eller andet y.

Figur 9.4: Biler på en landevej: Fraktildiagram over ventetiderne mellem på hinanden følgende biler, tegnet på grundlag af de rå data (Tabel 9.4). Den indlagte linie svarer til eksponentialfordelingen med parameter $\widehat{\lambda} = 0.0736$ sek^{-1}.



Kommentarer til den observerede ventetidsfordeling

Når man betragter histogrammet over fordelingen af ventetiden mellem på hinanden følgende biler (Figur 9.2) falder det i øjnene, at der er for mange store observationer – der er en klump mellem 30 og 45 sekunder som ikke harmonerer med eksponentialfordelings-antagelsen.

Fraktildiagrammet Figur 9.4 viser det samme, hvilket her ytrer sig derved, at punkterne et langt stykke ligger *over* linien og derefter *under*.

Fraktildiagrammet viser også, hvad der behændigt er skjult i histogrammet, at der er for få helt små observationer, – faktisk er den mindste observation 0.9 sek (jf. Tabel 9.3), og det er ikke særlig foreneligt med eksponentialfordelingsantagelsen. Lad os nemlig prøve at regne ud, hvad sandsynligheden er for at man blandt 64 eksponentialfordelte observationer ikke får nogen som er mindre end et bestemt tal y (= 0.9 sek). Sandsynligheden for, at en stokastisk variabel Y som er eksponentialfordelt med parameter λ er større end y, er $\mathrm{P}(Y > y) = \exp(-\lambda y)$; sandsynligheden for at n (= 64) uafhængige identisk fordelte sådanne Y-er alle er større end y er da

$$\mathrm{P}(Y > y)^n = \exp(-n\lambda y)$$

der i vores tilfælde kan estimeres til

$$\begin{aligned} \exp(-n\widehat{\lambda}y) &= \exp(-64 \times 0.0736 \times 0.9) \\ &= \exp(-4.239) \end{aligned}$$

hvilket er knap 1.5%. Den mindste observation er altså så stor at der kun er halvanden procents chance for at den var endnu større. Noget tyder altså på, at der altid er en vis mindste registreret ventetid mellem to biler, enten af tekniske grunde eller fordi bilerne faktisk altid kører med en vis afstand.

Kapitel 10

Normalfordelingen

Man har meget ofte brug for en type sandsynlighedsfordelinger der kan beskrive hvordan målinger el.lgn. varierer tilfældigt omkring et bestemt niveau, når det skal være sådan at de faktiske værdier lige så godt tilfældigvis kan være lidt *over* som de tilfældigvis kan være lidt *under* det teoretisk rigtige niveau.

For at kunne finde frem tilsådanne fordelinger må vi først formulere problemet mere præcist. Fordelingerne skal benyttes i statistiske modeller til beskrivelse af den tilfældige variation af målinger af længder, masser, koncentrationer osv., altsammen størrelser der måles på en *kontinuert* skala. Første punkt i problempræciseringen er defor:

1. *Der søges en type kontinuerte fordelinger.*

Fordelingerne skal beskrive den tilfældige variation omkring et vist niveau. Dette niveau skal indgå som en parameter μ, så modelfunktionen skal være en funktion af en observationsvariabel x og en parametervariabel μ:

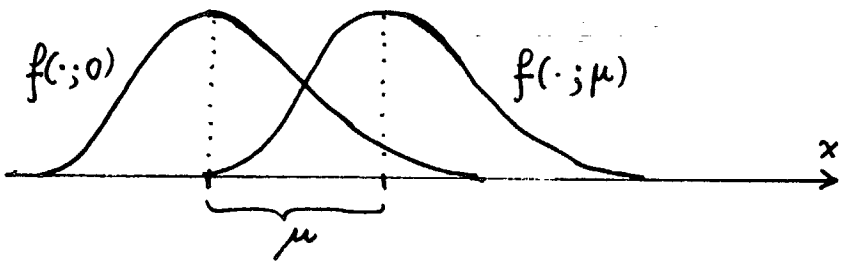
2. *Modelfunktionen er* $f(x;\mu)$.

Parameteren μ skal beskrive *hvor* på tallinien fordelingen er beliggende, og en ændring af parameterværdien skal svare til en forskydning af sandsynlighedsfordelingen hen ad tallinien (uden at fordelingens form i øvrigt ændres). Mere præcist vil vi antage at

3. Fordelingen svarende til parameterværdien μ fås ved at forskyde fordelingen svarende til parameterværdien 0 stykket μ, dvs.

$$f(x;\mu) \;=\; f(x-\mu;0)\;,$$

hvor μ i princippet kan antage alle mulige værdier.



Betingelse 3. udtrykker man også på den måde, at μ skal være en *positionsparameter*.

Betingelserne 1, 2 og 3 er ikke nok til at fastlægge fordelingen, så man er nødt til at stille nogle flere krav. Vi vil stille en *statistisk betingelse*, en betingelse der handler om, hvordan man skal analysere data fra den søgte fordeling. Da parameteren μ skal beskrive det niveau hvoromkring observationerne fordeler sig, kan man synes at det må være rimeligt at den ukendte parameter μ skal estimeres ved *gennemsnittet af observationerne*. Da det er et gennemgående princip at man altid skal benytte *maksimaliseringsestimater*, vil vi derfor stille dette krav:

4. Maksimaliseringsestimatet over μ skal være gennemsnittet af observationerne.

Vi skal nu se hvad der følger af disse betingelser.

Udledning af normalfordelingen

Vi skal i dette afsnit vise, at normalfordelingen er svar*et* på ønsket om en type kontinuerte fordelinger på den reelle akse, således at de er parametriseret med en positionsparameter og således at maksimaliseringsestimatet for positionsparameteren er gennemsnittet af observationerne. Det går for sig således:

1. Modelfunktionen svarende til et forsøg med én observation kaldes som nævnt $f(x;\mu)$. Modelfunktionen svarende til et forsøg med n observationer $(x_1, x_2, \ldots, x_n)$ er derfor $\prod_{i=1}^{n} f(x_i;\mu)$, så *likelihood-funktionen* er

$$L(\mu) = \prod_{i=1}^{n} f(x_i;\mu) . \tag{10.1}$$

2. Da der skal være tale om en *positionsparameter*, må der gælder at

$$\begin{aligned} f(x;\mu) &= f(x-\mu;0) \\ &= f_0(x-\mu) , \end{aligned}$$

hvor f_0 blot er en anden betegnelse for $f(\,\cdot\,;0)$. Likelihoodfunktionen (10.1) kan derfor skrives som

$$L(\mu) = \prod_{i=1}^{n} f_0(x-\mu) ,$$

og *log-likelihoodfunktionen* er tilsvarende

$$\ln L(\mu) = \sum_{i=1}^{n} \ln f_0(x-\mu) . \tag{10.2}$$

3. Ifølge vore antagelser skal (10.2) antage sin maksimale værdi når μ har værdien $\widehat{\mu} = \overline{x}$. Vi går stiltiende ud fra, at f_0 og dermed også $\ln L$ er en pæn differentiabel funktion. I så fald gælder, at da $\ln L$ skal have maksimum når μ har værdien $\overline{x}$, så må differentialkvotienten $(\ln L)'(\mu)$ være 0 når μ har værdien $\overline{x}$: $(\ln L)'(\overline{x}) = 0$.

4. Af (10.2) fås

$$\begin{aligned} (\ln L)'(\mu) &= \sum_{i=1}^{n} -(\ln f_0)'(x-\mu) \\ &= \sum_{i=1}^{n} g(x-\mu) , \end{aligned}$$

hvor g er en kort betegnelse for $-(\ln f_0)'$. Betingelsen om, at maksimaliseringsestimatet skal være gennemsnittet $\overline{x}$, betyder derfor at funktionen g skal opfylde betingelsen

$$\sum_{i=1}^{n} g(x-\overline{x}) = 0 . \tag{10.3}$$

5. Fidusen er nu at (10.3) skal gælde for *alle* valg af $x_1, x_2, \ldots, x_n$, og ved at indsætte nogle tilpas snedigt valgte x-er kan man få at vide hvordan funktionen g nødvendigvis må se ud.

   (a) Ved at vælge $n = 2$ og $x_2 = -x_1 = y$ (hvorved $\overline{x} = 0$) fås af (10.3) at $g(-y) + g(y) = 0$, dvs.

   $$g(-y) = -g(y) \tag{10.4}$$

   for vilkårligt y.

   (b) Ved at vælge $n = k + 1$ og lade de k første x-er være ens og lade gennemsnittet være 0, mere præcist $x_1 = x_2 = \ldots = x_k = -y$ og $x_{k+1} = ky$, fås at $k\,g(-y) + g(k\,y) = 0$, der ved brug af (10.4) kan formuleres som

   $$g(k\,y) = k\,g(y)\,, \tag{10.5}$$

   gældende for vilkårligt y og $k = 1, 2, 3, \ldots$

   (c) I formel (10.5) kan vi vælge $y = j/k$ hvor j og k er heltal. Derved fås at $g(j) = k\,g(j/k)$, dvs. at $g(j/k) = \frac{1}{k}g(j)$.
   Men vi kan også vælge $y = 1$ og $k = j$ i (10.5) og derved få at $g(j) = j\,g(1)$. Alt i alt er dermed $g(j/k) = \frac{j}{k}g(1)$, hvilket vi formulerer således:

   $$\begin{aligned} g(y) &= y\,g(1) \\ &= g(1)\,y \end{aligned} \tag{10.6}$$

   for alle rationale y.

   Medmindre g skal være en ganske overordentlig usædvanlig funktion så er det sådan, at når (10.6) gælder for alle *rationale* tal y, så gælder (10.6) også for alle *reelle* tal y. Vi vil gå ud fra at (10.6) gælder for alle y, og vi er altså nået frem til, at funktionen g er en ganske almindelig lineær funktion:

   $$g(x) = c\,x$$

   for en passende valgt konstant c.

6. Da g blot var funktionen $-(\ln f_0)'$, kan vi dernæst finde f_0: Hvis $-(\ln f_0)'(x) = c\,x$, så er

   $$\ln f_0(x) = -\tfrac{1}{2}c\,x^2 + \text{ konstant}\,,$$

dvs.

$$f_0(x) \;=\; \text{konstant} \times \exp(-\tfrac{1}{2}c\,x^2)\;.$$

7. Denne funktion f_0 skal være en sandsynlighedstæthed, hvilket vil sige at den skal være ikke-negativ og integrere til 1, dvs. $\int_{-\infty}^{+\infty} f_0(x)dx = 1$. For at dette sidste skal kunne lade sig gøre må konstanten c nødvendigvis være positiv; traditionen tro omdøber vi c til $1/\sigma^2$, så at tæthedsfunktionen får udseendet

$$f_0(x) \;=\; \text{konstant} \times \exp\left(-\tfrac{1}{2}\frac{x^2}{\sigma^2}\right)\;.$$

Den betingelse at f_0 skal integrere til 1 fastlægger "konstant"; man kan vise[1], at "konstanten" skal være $1/\sqrt{2\pi\sigma^2}$. Dermed har vi fundet at

$$f_0(x) \;=\; \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\tfrac{1}{2}\frac{x^2}{\sigma^2}\right)$$

og dermed

$$\begin{aligned} f(x;\mu) \;&=\; f_0(x-\mu) \\ &=\; \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\tfrac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)\;. \end{aligned} \qquad (10.7)$$

Det oprindelige problem bestod i at finde en type fordelinger hvor der indgik en *positionsparameter* μ. I den fundne løsning (10.7) optræder imidlertid også en størrelse σ^2 der er kommet ind i billedet som en integrationskonstant. Størrelsen σ^2 udnævner vi til en *parameter* og samtidig omdøbes $f(x;\mu)$ til $f(x;\mu,\sigma^2)$:

$$f(x;\mu,\sigma^2) \;=\; \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\tfrac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)\;.$$

I *Grundbegreber i Sandsynlighedsregningen* er der gjort rede for, at for ethvert valg af $\mu \in]-\infty,+\infty[$ og $\sigma^2 > 0$ er (10.7) en sandsynlighedstæthedsfunktion; denne tæthedsfunktion er tætheden for *normalfordelingen med positionsparameter μ og kvadratisk skalaparameter σ^2*, kort $\mathcal{N}(\mu\,,\,\sigma^2)$.

[1] Se f.eks. Kapitel 4 i *Grundbegreber i Sandsynlighedsregningen*.

Resultatet af ovenstående udledninger er således, at *hvis* vi er på jagt efter en type kontinuerte sandsynlighedsfordelinger hvor der optræder en positionsparameter, og *hvis* vi forlanger at denne positionsparameter skal estimeres ved gennemsnittet af observationerne, så er normalfordelinger[2] den eneste type fordelinger der kan komme på tale. – Strengt taget har vi ikke vist at normalfordelingerne faktisk har den ønskede egenskab, men det kommer i det følgende.

Egenskaber ved normalfordelingen

Her gives en oversigt over forskellige egenskaber o.lgn. ved normalfordelingen[3].

- Normalfordelingen med parametre μ og σ^2, kort $\mathcal{N}(\mu\,,\,\sigma^2)$-fordelingen, er den sandsynlighedsfordeling på den reelle talakse $\mathbf{R}$ som har tæthedsfunktionen

$$f(x;\mu,\sigma^2) \;=\; \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\tfrac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)\,.$$

  Her kan parameteren μ være et vilkårligt reelt tal og parameteren σ^2 et vilkårligt positivt tal.

- Parameteren μ er en *positionsparameter*, dvs. hvis X er $\mathcal{N}(\mu\,,\,\sigma^2)$-fordelt og a en konstant, så vil $a+X$ være $\mathcal{N}(a+\mu\,,\,\sigma^2)$-fordelt. Desuden er μ *middelværdien* i $\mathcal{N}(\mu\,,\,\sigma^2)$-fordelingen. Endvidere er μ *medianen*[4] i $\mathcal{N}(\mu\,,\,\sigma^2)$-fordelingen.

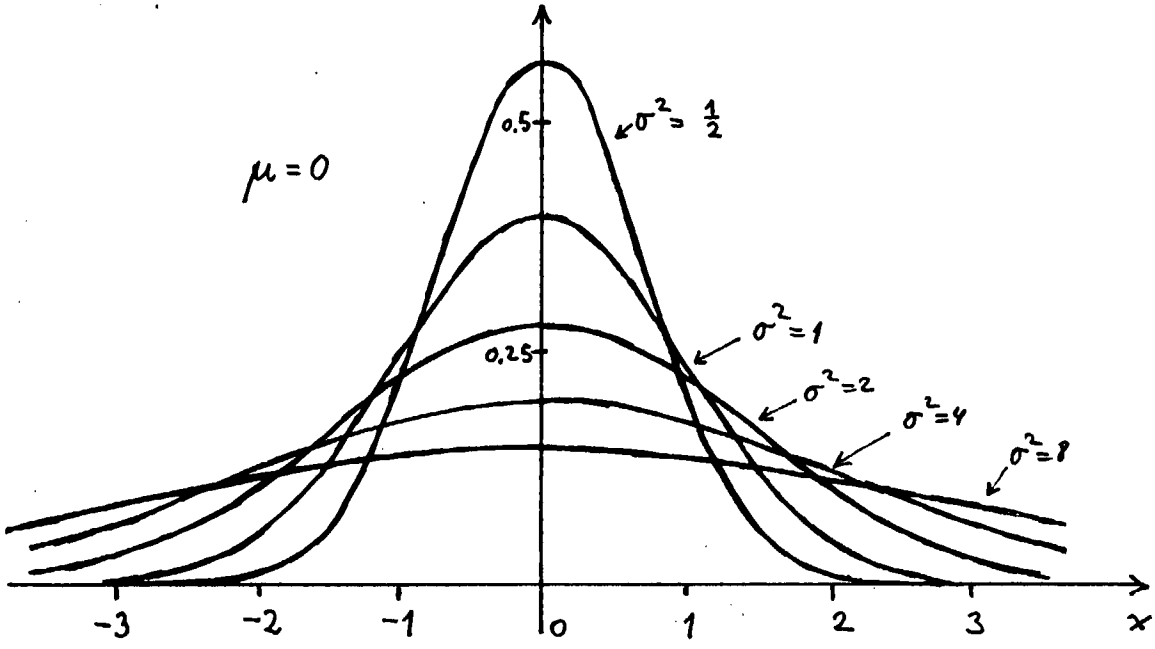
- Parameteren σ^2 er en *kvadratisk skalaparameter*, dvs. hvis X er $\mathcal{N}(0\,,\,\sigma^2)$-fordelt og b en konstant, så vil bX være $\mathcal{N}(0\,,\,b^2\sigma^2)$-fordelt. Desuden er σ^2 *variansen* i $\mathcal{N}(\mu\,,\,\sigma^2)$-fordelingen, og dermed er σ standardafvigelsen $\mathcal{N}(\mu\,,\,\sigma^2)$-fordelingen. Undertiden

[2]Normalfordelinger kaldes også Gauß-fordelinger. Gauß (1777 - 1855) benyttede bl.a. normalfordelinger til at beskrive astronomiske målingers tilfældige afvigelser omkring den sande værdi. I værket *Theoria Motus Corporum Coelestium in Sectionibus Conicus Arbientium* (dvs. Teori om de himmelske legemers bevægelser i keglesnit omkring solen) argumenterede han for normalfordelingen på en måde der meget ligner den der er benyttet her.

[3]Der bliver ikke givet beviser for de forskellige påstande; en del af påstandene er bevist i *Grundbegreber i Sandsynlighedsregningen*.

[4]dvs. 50%-fraktilen

Figur 10.1: Nogle normalfordelingstætheder



kaldes $1/\sigma^2$ for *præcisionen* i fordelingen, fordi $1/\sigma^2$ er et udtryk for hvor snævert fordelingen er koncentreret om sin middelværdi.

- Hvis X er $\mathcal{N}(\mu\,,\,\sigma^2)$-fordelt, så vil $a+bX$ være $\mathcal{N}(a+b\mu\,,\,b^2\sigma^2)$-fordelt; her betegner a og b konstanter.

- Den *normerede normale fordeling* er $\mathcal{N}(0\,,\,1)$-fordelingen. Dens tæthed betegnes som oftest ϕ:

$$\phi(x) \;=\; \frac{1}{\sqrt{2\pi}}\exp(-\tfrac{1}{2}x^2)$$

og dens kumulerede fordelingsfunktion betegnes tilsvarende Φ, dvs. $\Phi(u)$ er sandsynligheden for at en $\mathcal{N}(0\,,\,1)$-variabel er mindre end eller lig u:

$$\begin{aligned}\Phi(u) \;&=\; \int_{-\infty}^{u}\phi(x)dx \\ &=\; \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{u}\exp(-\tfrac{1}{2}x^2)dx\;.\end{aligned}$$

- En $\mathcal{N}(\mu\,,\,\sigma^2)$-variabel har tæthedsfunktion

$$x \;\mapsto\; \frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})$$

og kumuleret fordelingsfunktion

$$x \mapsto \Phi(\frac{x-\mu}{\sigma}) .$$

- Hvis α er et tal mellem 0 og 1 så har ligningen $\Phi(u) = \alpha$ præcis én løsning u_α, hvor u_α betegner *α-fraktilen* i den normerede normale fordeling. Ved at lægge fem til fraktilerne fås de såkaldte *probits*[5]:

$$\text{probit}(\alpha) \;=\; u_\alpha + 5 .$$

I statistiske tabelværker findes tabeller over $\Phi(u)$ og over fraktilerne u_α eller $u_\alpha + 5$.

Enstikprøveproblemet i normalfordelingen

Normalfordelingen blev indført i forbindelse med jagten på en fordeling hvor positionsparameteren estimeres ved gennemsnittet af observationerne. Vi mangler at gøre rede for, at normalfordelingen faktisk *har* den ønskede egenskab. Vi skal nu se på, hvordan man skal estimere μ og σ^2.

Enstikprøveproblemet i normalfordelingen er den situation der består i at man har en stikprøve, altså et antal uafhængige observationer $y_1, y_2, \ldots, y_n$, fra en $\mathcal{N}(\mu, \sigma^2)$-fordeling. Parametrene μ og σ^2 er ukendte, og problemet er at bestemme estimater over dem og måske teste statistiske hypoteser om dem. En anden side af enstikprøveproblemet i normalfordelingen er *modelkontrolproblemet*, dvs. hvordan vurderer man rimeligheden i at gå ud fra, at observationerne nu også *er* normalfordelte. Eksempel 10.1 kan tjene som eksempel til "enstikprøveproblemet i normalfordelingen".

[5]probit = **pro**bability **unit**

Tabel 10.1: Newcombs bestemmelser af lysets passagetid af en strækning på 7442 m.
Tabelværdierne $\times 10^{-3} + 24.8$ er passagetiden i 10^{-6} sek.

28	26	33	24	34	-44
27	16	40	-2	29	22
24	21	25	30	23	29
31	19	24	20	36	32
36	28	25	21	28	29
37	25	28	26	30	32
36	26	30	22	36	23
27	27	28	27	31	27
26	33	26	32	32	24
39	28	24	25	32	25
29	27	28	29	16	23

Eksempel 10.1. ***Lysets hastighed***

I årene 1879-82 foretog den amerikanske fysiker A.A. Michelson og den amerikanske matematiker og astronom S. Newcomb en række (efter den tids forhold) temmelig nøjagtige bestemmelser af lysets hastighed i luft. Deres metoder var baseret på Foucault's idé med at sende en lysstråle fra et hurtigt roterende spejl hen på et fjernt fast spejl som returnerer lysstrålen til det roterende, hvor man måler dens vinkelforskydning i forhold til den oprindelige lysstråle. Hvis man kender rotationshastigheden samt afstanden mellem spejlene, kan man derved bestemme lyshastigheden.

I Tabel 10.1 er vist resultaterne af de 66 målinger som Newcomb foretog i perioden 24. juli til 5. september 1882 i Washington, D.C. I Newcombs opstilling var der 3721 m mellem det roterende spejl, der var placeret i Fort Myer på vestbredden af Potomac-floden, og det faste spejl, der var anbragt på George Washington-monumentets fundament. De størrelser som Newcomb rapporterer er lysets passagetid, *altså den tid som det er om at tilbagelægge den pågældende distance.*

Af de 66 værdier i Tabel 10.1 skiller to sig ud, nemlig -44 og -2, der synes at være 'outliers', altså tal der tilsyneladende ligger

for langt væk fra flertallet af observationerne. Det er altid et vanskeligt spørgsmål at afgøre, om det er forsvarligt at se bort fra sådanne 'outliere'. I analysen af tallene i Tabel 10.1 vil vi dog her vælge at se bort fra de to nævnte observationer, således at vi kun har at gøre med 64 observationer. □

I den generelle situation foreligger der størrelser $y_1, y_2, \ldots, y_n$ der antages at være observerede værdier af stokastiske variable $Y_1, Y_2, \ldots, Y_n$ som er uafhængige identisk $\mathcal{N}(\mu\,,\,\sigma^2)$-fordelte; her er μ og σ^2 ukendte parametre. *Modelfunktionen* er

$$f(y_1, y_2, \ldots, y_n; \mu, \sigma^2) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\tfrac{1}{2}\frac{(y_j-\mu)^2}{\sigma^2}\right)$$

$$= (2\pi)^{-n/2}(\sigma^2)^{-n/2} \exp\left(-\tfrac{1}{2}\frac{\sum_{j=1}^{n}(y_j-\mu)^2}{\sigma^2}\right). \qquad (10.8)$$

Likelihoodfunktionen svarende til observationerne $y_1, y_2, \ldots, y_n$ er derfor

$$L(\mu, \sigma^2) = \text{konstant} \times (\sigma^2)^{-n/2} \exp\left(-\tfrac{1}{2}\frac{\sum_{j=1}^{n}(y_j-\mu)^2}{\sigma^2}\right). \qquad (10.9)$$

Estimation af μ og σ^2

Vi vil bestemme maksimaliseringsestimaterne for μ og σ^2. Af udtrykket (10.9) for likelihoodfunktionen ser vi, at ligegyldigt hvilken værdi σ^2 måtte have, så er den bedste μ-værdi, altså den μ-værdi som maksimaliserer $\mu \mapsto L(\mu, \sigma^2)$, den værdi som *minimaliserer* kvadratsummen

$$\sum_{j=1}^{n}(y_j-\mu)^2\,.$$

Ved at benytte formlen for kvadratet på en to-leddet størrelse kan denne kvadratsum omskrives således, hvor $\overline{y}$ betegner gennemsnittet af y-erne:

$$\begin{aligned}
&\sum_{j=1}^{n}(y_j-\mu)^2 \\
&= \sum_{j=1}^{n}\Big((y_j-\overline{y})+(\overline{y}-\mu)\Big)^2 \\
&= \sum_{j=1}^{n}\Big((y_j-\overline{y})^2+2(y_j-\overline{y})(\overline{y}-\mu)+(\overline{y}-\mu)^2\Big) \\
&= \sum_{j=1}^{n}(y_j-\overline{y})^2+\sum_{j=1}^{n}2(y_j-\overline{y})(\overline{y}-\mu)+\sum_{j=1}^{n}(\overline{y}-\mu)^2 \\
&= \sum_{j=1}^{n}(y_j-\overline{y})^2+2(\overline{y}-\mu)\sum_{j=1}^{n}(y_j-\overline{y})+n\,(\overline{y}-\mu)^2 \\
&= \sum_{j=1}^{n}(y_j-\overline{y})^2+n(\overline{y}-\mu)^2\,,
\end{aligned}$$

altså

$$\sum_{j=1}^{n}(y_j-\mu)^2 \;=\; \sum_{j=1}^{n}(y_j-\overline{y})^2+n(\overline{y}-\mu)^2\,. \qquad (10.10)$$

Heraf ses at kvadratsummen er mindst netop når μ er lig med $\overline{y}$. Derfor *er* maksimaliseringsestimatet for μ faktisk gennemsnittet af observationerne,

$$\widehat{\mu} \;=\; \overline{y}\,,$$

således som det jo også var tanken at det skulle være.

Herefter kan man bestemme maksimaliseringsestimatet for σ^2 som maksimumspunktet for funktionen

$$\sigma^2 \;\mapsto\; L(\overline{y},\sigma^2)\,. \qquad (10.11)$$

Man vil finde at (10.11) antager sit maksimum når σ^2 har værdien

$$\widehat{\sigma^2} \;=\; \frac{1}{n}\sum_{j=1}^{n}(y_j-\overline{y})^2\,.$$

Imidlertid benytter man som regel *ikke* dette skøn over σ^2, men derimod

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_j - \overline{y})^2 ,$$

hvor divisoren $n-1$ i denne forbindelse kaldes for *antallet af frihedsgrader* for variansskønnet s^2. I taleksemplet finder man (idet vi går ud fra, at de 64 passagetider kan betragtes som 64 observationer fra en og samme normalfordeling) at middelværdien estimeres til $\overline{y} = 27.75$ og den estimerede varians er $s^2 = 25.8$ med 63 frihedsgrader[6].

Hvorfor benyttes s^2?

De to parametre μ og σ^2 i normalfordelingen opfattes sædvanligvis ikke som værende ligestillede. Man plejer at tænke på middelværdiparameteren (μ) som den primære, da den jo beskriver *den systematiske variation*, nemlig det *niveau* hvorom observationerne fordeler sig, hvorimod variansparameteren σ^2, der "kun" beskriver den tilfældige variation, kommer i anden række. Som en konsekvens heraf kan man mene, at man ikke skal estimere de to parametre samtidigt, men at man *først* skal estimere μ og *dernæst* σ^2. Man skal endda til estimationen af σ^2 kun benytte det der er tilbage af (informationen i) talmaterialet efter at man først har estimeret μ.

Hvis der f.eks. foreligger de fem observationer 3.2, 5.7, 2.1, 7.4, 3.1, som tænkes at stamme fra en $\mathcal{N}(\mu\,,\,\sigma^2)$-fordeling, så estimeres først den "væsentlige" parameter μ ved gennemsnittet $(3.2 + 5.7 + 2.1 + 7.4 + 3.1)/5 = 21.5/5 = 4.3$. Dernæst skal man estimere σ^2, der skal beskrive den tilfældige variation omkring niveauet 4.3. Da det nu i en vis forstand er *givet* at de fem værdier skal have gennemsnit 4.3, dvs. at de

[6]Dvs. at passagetidens middelværdi estimeres til

$$(27.75 \times 10^{-3} + 24.8) \times 10^{-6} \text{ sek} = 24.828 \times 10^{-6} \text{ sek}$$

og variansen på passagetiden estimeres til

$$25.8 \times (10^{-3} \times 10^{-6} \text{ sek})^2 = 25.8 \times 10^{-6}(10^{-6} \text{ sek})^2$$

med 63 frihedsgrader, dvs. standardafvigelsen estimeres til

$$\sqrt{25.8 \times 10^{-6}}10^{-6} \text{ sek} = 0.005 \times 10^{-6} \text{ sek} .$$

fem afvigelser fra gennemsnittet[7] skal summere til 0, så er der på sin vis kun *fire* forskellige afvigelser. Når man skal estimere variansen (= den forventede kvadratiske afvigelse af en observation fra middelværdien) bliver det derfor som summen af de kvadratiske afvigelser divideret med fire:

$$\left((3.2-4.3)^2+(5.7-4.3)^2+(2.1-4.3)^2+(7.4-4.3)^2+(3.1-4.3)^2\right)/4$$

$$\begin{aligned} &= \left((-1.1)^2+1.4^2+(-2.2)^2+3.1^2+(-1.2)^2\right)/4 \\ &= 19.08/4 \\ &= 4.77\ . \end{aligned}$$

Man siger at der er fire *frihedsgrader* fordi når det er fixeret at de fem observationer skal have et bestemt gennemsnit (4.3), så kan man vælge fire af de fem afvigelser fra gennemsnittet frit.

Ovenstående argument for at dividere summen af de kvadratiske afvigelser med $n-1$ i stedet for med n kan jo godt siges at være noget løst og "høker-agtigt". Der findes også mere "matematiske" ræsonnementer; her skitseres først ét som er en del i familie med det løse. Det forhold, at variansparameteren σ^2 tænkes at spille en underordnet rolle i forhold til middelværdiparameteren μ og at dette skal afspejles i den måde, som parametrene skal estimeres på, kan formaliseres på følgende måde: Man skal først estimere μ på sædvanlig måde, men dernæst skal man estimere σ^2 i den *betingede model* hvor man betinger med estimatet for μ, altså med $\overline{y}$. Estimatet over σ^2 skal være maximum likelihood estimatet, men man skal vel at mærke benytte likelihoodfunktionen svarende til *den betingede fordeling af* $Y_1, Y_2, \ldots, Y_n$ *givet at* $\overline{Y}$ *er lig med* $\overline{y}$. Hvis det skal gå matematisk nogenlunde korrekt til er det ikke noget simpelt problem at bestemme denne betingede fordeling – det handler jo om kontinuerte fordelinger; men hvis man i al naivitet regner med at resultatet vil ligne formel (1.3) i *Grundbegreber i Sandsynlighedsregningen*, blot med tætheder i stedet for sandsynligheder, så skulle den betingede tæthedsfunktion være

$$\frac{\text{tæthedsfunktionen for } Y_1, Y_2, \ldots, Y_n}{\text{tæthedsfunktionen for } \overline{Y}} \qquad (10.12)$$

[7] de fem afvigelser er -1.1, 1.4, -2.2, 3.1, -1.2

Da $Y_1, Y_2, \ldots, Y_n$ er uafhængige $\mathcal{N}(\mu\,,\,\sigma^2)$-variable, vil gennemsnittet $\overline{Y}$ være $\mathcal{N}(\mu\,,\,\sigma^2/n)$-fordelt[8]. Derfor kan (10.12) omskrives til (jf. (10.8))

$$\frac{(2\pi)^{-n/2}\,(\sigma^2)^{-n/2} \times \exp\left(-\frac{1}{2}\dfrac{\sum_{j=1}^{n}(y_j-\mu)^2}{\sigma^2}\right)}{\dfrac{1}{\sqrt{2\pi\sigma^2/n}}\exp\left(-\frac{1}{2}\dfrac{(\overline{y}-\mu)^2}{\sigma^2/n}\right)}$$

$$= \text{konstant} \times (\sigma^2)^{-\frac{n-1}{2}} \exp\left(-\frac{1}{2}\frac{\sum_{j=1}^{n}(y_j-\overline{y})^2}{\sigma^2}\right).$$

Opfattet som funktion af σ^2 skulle dette så være den betingede likelihoodfunktion[9], altså den likelihoodfunktion der skal benyttes ved estimation af σ^2. Den betingede likelihoodfunktion er en ganske almindelig funktion af én variabel σ^2, og man kan uden videre benytte den almindelige metode til bestemmelse af maksimumspunkter, dvs. bestem de punkter hvor funktionens[10] første afledede er 0 og den anden afledede negativ. Man finder da, at funktionen antager sit maksimum i ét punkt, nemlig når σ^2 har værdien s^2. Der gælder altså derfor rent faktisk at

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_j-\overline{y})^2$$

er et maximum likelihood estimat over σ^2.

Et helt andet "argument" for at benytte s^2 og ikke $\widehat{\sigma^2}$ som skøn over σ^2 er, at s^2 i modsætning til $\widehat{\sigma^2}$ er et såkaldt *centralt* skøn over σ^2. At s^2 er et centralt skøn over σ^2 betyder, at middelværdien af den stokastiske variabel s^2 er lig σ^2, altså $\mathrm{E}\,s^2 = \sigma^2$.

[8] Jf. Afsnit 4.4 i *Grundbegreber i Sandsynlighedsregningen*.

[9] Bemærk at μ meget bekvemt er forsvundet ud af billedet.

[10] Det kan være bekvemt at betragte *log*-likelihoodfunktionen.

Test af hypotese om middelværdien

Man er undertiden interesseret i at undersøge, om de foreliggende data er forenelige med en antagelse om, at den teoretiske middelværdi μ har en bestemt værdi (f.eks. 0). Mere formelt ønsker man at teste den statistiske hypotese $H_0 : \mu = \mu_0$, hvor μ_0 er et kendt tal.

Hypoteser om parametre i normalfordelinger testes principielt på samme måde som alle andre statistiske hypoteser, nemlig ved brug af et kvotienttest der sammenligner likelihoodfunktionens maksimale værdi under hypotesen med den maksimale værdi overhovedet under den givne model. Den maksimale værdi overhovedet af likelihoodfunktionen (10.9) er $L(\overline{y}, \widehat{\sigma^2})$. Under H_0 er likelihoodfunktionen

$$L_0(\sigma^2) \;=\; L(\mu_0, \sigma^2)$$

og den antager sin maksimumsværdi når σ^2 er lig med

$$\widehat{\widehat{\sigma^2}} \;=\; \frac{1}{n}\sum_{j=1}^{n}(y_j - \mu_0)^2 \,.$$

Kvotienttesstørrelsen bliver dermed

$$\begin{aligned} Q \;&=\; \frac{L_0(\widehat{\widehat{\sigma^2}})}{L(\overline{y}, \widehat{\sigma^2})} \\ &=\; \left(\frac{\widehat{\widehat{\sigma^2}}}{\widehat{\sigma^2}}\right)^{-n/2} \exp\left(-\tfrac{1}{2}\left(\frac{\sum_{j=1}^{n}(y_j-\mu_0)^2}{\widehat{\widehat{\sigma^2}}} - \frac{\sum_{j=1}^{n}(y_j-\overline{y})^2}{\widehat{\sigma^2}}\right)\right) \\ &=\; \left(\frac{\sum_{j=1}^{n}(y_j-\mu_0)^2}{\sum_{j=1}^{n}(y_j-\overline{y})^2}\right)^{-n/2} \exp(-\tfrac{1}{2}(n-n)) \\ &=\; \left(\frac{\sum_{j=1}^{n}(y_j-\mu_0)^2}{\sum_{j=1}^{n}(y_j-\overline{y})^2}\right)^{-n/2} . \end{aligned}$$

Her omskrives kvadratsummen i tælleren ved hjælp af (10.10) (med μ erstattet af μ_0) og man får

$$\begin{aligned} Q &= \left(\frac{\sum_{j=1}^{n}(y_j - \overline{y})^2 + n(\overline{y} - \mu_0)^2}{\sum_{j=1}^{n}(y_j - \overline{y})^2} \right)^{-n/2} \\ &= \left(1 + \frac{n(\overline{y} - \mu_0)^2}{\sum_{j=1}^{n}(y_j - \overline{y})^2} \right)^{-n/2} \\ &= \left(1 + \frac{n(\overline{y} - \mu_0)^2}{(n-1)s^2} \right)^{-n/2} \\ &= \left(1 + \frac{1}{n-1} \left(\frac{\overline{y} - \mu_0}{\sqrt{s^2/n}} \right)^2 \right)^{-n/2} . \end{aligned}$$

Størrelsen $\dfrac{\overline{y} - \mu_0}{\sqrt{s^2/n}}$ plejer man at betegne t:

$$t = \frac{\overline{y} - \mu_0}{\sqrt{s^2/n}} , \tag{10.13}$$

og med denne betegnelse har vi at

$$Q = \left(1 + \frac{t^2}{n-1} \right)^{-n/2} . \tag{10.14}$$

Nu er det jo sådan at *små* værdier af Q tyder på at hypotesen H_0 *ikke* er forenelig med data, og (10.14) viser, at små Q-værdier er ensbetydende med t-værdier langt fra 0, dvs. med store $|t|$-værdier. Man kan altså benytte t som teststørrelse i stedet for Q, hvilket er praktisk da t er lettere at regne ud end Q er. t-teststørrelsen kaldes undertiden for *'Student's t*, fordi W.S. Gosset, der skrev den første artikel om t-testet (i 1908), skrev under pseudonymet 'Student'.

Bemærk at t-teststørrelsen også ud fra en umiddelbar betragtning forekommer at være en fornuftig teststørrelse, idet den måler afvigelsen

$\overline{y} - \mu_0$ mellem den observerede og den teoretiske middelværdi i forhold til skønnet $\sqrt{s^2/n}$ over middelfejlen på $\overline{y}$, dvs. standardafvigelsen på $\overline{y}$.

Når man har fundet værdien af teststørrelsen t, går næste skridt i testproceduren ud på at bestemme *testsandsynligheden*, altså sandsynligheden for at få en mere ekstrem værdi af teststørrelsen end den faktisk opnåede, forudsat at hypotesen H_0 er rigtig. En matematisk sætning fortæller, at når H_0 er rigtig, så følger t-størrelsen en ganske bestemt slags fordeling[11], nemlig en såkaldt *t-fordeling med $f = n - 1$ frihedsgrader*. Frihedsgradsantallet kommer fra frihedsgradsantallet for variansskønnet s^2 i nævneren. I statistiske tabelværker kan man finde tabeller over fraktiler i t-fordelingen, og ved hjælp af sådanne tabeller er det ganske let at bestemme (omtrentlige) værdier af t-testsandsynligheder. Man skal dog være opmærksom på en enkelt ting, nemlig at en "mere ekstrem t-værdi" som regel vil sige en t-værdi således at $|t| > |t_{\text{obs}}|$, dvs.

$$t > |t_{\text{obs}}| \quad \text{eller} \quad t < -|t_{\text{obs}}| \; ;$$

man vil altså forkaste hypotesen både hvis t_{obs} er meget stor og hvis den er meget lille. (Et sådant test kaldes et *tosidet test*, i modsætning til et *ensidet test* der regner med at de "ekstreme" afvigelser kun kan være til den ene side, f.eks. den positive, så at man kun forkaster hvis t_{obs} er meget stor.) Der gælder at t-fordelingen er symmetrisk omkring 0, hvilket indebærer at

$$\mathrm{P}\left(t > |t_{\text{obs}}|\right) \;=\; \mathrm{P}\left(t < -|t_{\text{obs}}|\right)$$

og dermed

$$\mathrm{P}\left(|t| > |t_{\text{obs}}|\right) \;=\; 2 \times \mathrm{P}\left(t > |t_{\text{obs}}|\right) \; .$$

Eksempel 10.2. *Lysets hastighed, fortsat*

I forbindelse med målingerne af lysets passagetid være interesseret i at undersøge, om resultaterne stemmer overens med det kendskab vi i dag har til værdien af lysets hastighed. Hvis vi går ud fra, at lysets hastighed er 2.998×10^8 *meter pr. sekund, så vil det være* $\tau_0 = 2.48232 \times 10^{-5}$ *sekunder om at tilbagelægge*

[11] der hverken afhænger af μ_0 eller af σ^2 hvilket er smart da vi jo ikke kender de nøjagtige værdier heraf.

strækningen på de 7442 m. *Størrelsen* τ_0 *svarer til en tabelværdi på* $((\tau_0 \times 10^6) - 24.8) \times 10^3 = 23.2$, *så det ville være interessant at undersøge, om de foreliggende data er forenelige med hypotesen om, at den ukendte middelværdi* μ *har værdien* $\mu_0 = 23.2$. *Derfor vil vi teste den statistiske hypotese* $H_0 : \mu = 23.2$. *Vi har fundet at* $\overline{y} = 27.75$ *og* $s^2 = 25.8$, *så t-teststørrelsen bliver*

$$\begin{aligned} t &= \frac{27.75 - 23.2}{\sqrt{25.8/64}} \\ &= \frac{4.55}{0.635} \\ &= 7.2 \, . \end{aligned}$$

Da der ikke er nogen grund til at tro at der kun skulle kunne forekomme afvigelser i én retning, skal testet være tosidet. Testsandsynligheden er altså sandsynligheden for at få t-værdier som enten er større end 7.2 eller mindre end -7.2. *Ved tabelopslag kan man finde, at i t-fordelingen med 63 frihedsgrader er 99.95%-fraktilen lidt over 3.4, dvs. der mindre end 0.05% sandsynlighed for at få en værdi som er større end 7.2, og testsandsynligheden er dermed mindre* $2 \times 0.05\% = 0.1\%$. *En så lille testsandsynlighed betyder at man må* forkaste *hypotesen. Newcomb's målinger af lysets passagetid stemmer altså* ikke *overens med hvad vi i dag ved om lysets hastighed. Vi ser at Newcomb's passagetider er en smule for store, og da den lyshastighed vi her har benyttet er lysets hastighed i vacuum, så kan noget af forklaringen være, at lyset bevæger sig en smule langsommere i luft end i vacuum.* □

Histogrammer og fraktildiagrammer for normalfordelte observationer

For at få en idé om modellens rimelighed vil man ofte i et "enstikprøveproblem i normalfordelingen" tegne histogrammer og fraktildiagrammer (jf. Kapitel 9).

Histogrammer over formodet normalfordelte observationer udarbejdes efter samme retningslinier som histogrammer i al almindelighed, jf. Kapitel 9, side 167. Som illustration vises her et histogram over de 64 passagetider for lyset. Når man skal udarbejde histogrammer

og fraktildiagrammer for data $y_1, y_2, \ldots, y_n$ er det som oftest hensigtsmæssigt først at danne *de ordnede observationer* $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$. De ordnede observationer for lys-dataene fremgår af Tabel 10.2. Et histogram svarende til intervallængden 2 er vist som Figur 10.2. Efter histogrammet at dømme kan vi med rimelighed anse observationerne for normalfordelte.

Når man skal tegne fraktildiagrammer for at undersøge om tal er normalfordelte kan man med fordel benytte det såkaldte *sandsynlighedspapir*[12]. På sandsynlighedspapiret er der to skalaer på ordinataksen. Den ene er en *probit-skala* som er ækvidistant og som går fra knap 2 til godt 8. Den anden er en (ikke-ækvidistant) *sandsynlighedsskala* med sandsynligheder i procent; denne skala går fra 0.05 til 99.95. Når man skal tegne fraktildiagrammet skal man først kende de ordnede observationer $y_{(i)}$ og værdierne af den empiriske fordelingsfunktion $\widehat{F}$ i punkterne $y_{(i)}$. Der gælder (jf. (9.3) side 170) at $\widehat{F}(y_{(i)})$ er lig antal observationer $< y_{(i)}$ plus halvdelen af antal observationer $= y_{(i)}$, det hele divideret med n. På sandsynlighedspapiret afsætter man punkterne $(y_{(i)}, \widehat{F}(y_{(i)}))$ idet man benytter sandsynlighedsskalaen på ordinataksen. Hvis tallene er normalfordelte skal punkterne fordele sig omkring en bestemt ret linie, der kan indtegnes ved at benytte probit-skalaen på ordinataksen, og lade linien gå gennem punkterne $(\overline{y}, 5)$, $(\overline{y} + s, 6)$, $(\overline{y} + 2s, 7)$, $(\overline{y} - s, 4)$, $(\overline{y} - 2s, 3)$ osv.

På Figur 10.3 er vist et fraktildiagram for lys-dataene. For ikke at skulle udregne og afsætte alt for mange punkter er dette fraktildiagram tegnet på grundlag af de grupperede observationer, jf. Tabel 10.2. Det vil sige at man som midtpunkter har benyttet intervalmidtpunkterne og som værdier af den empiriske fordelingsfunktion "antal obserationer til venstre for intervallet plus halvdelen af antal observationer i intervaller divideret med 64", se Tabel 10.3.

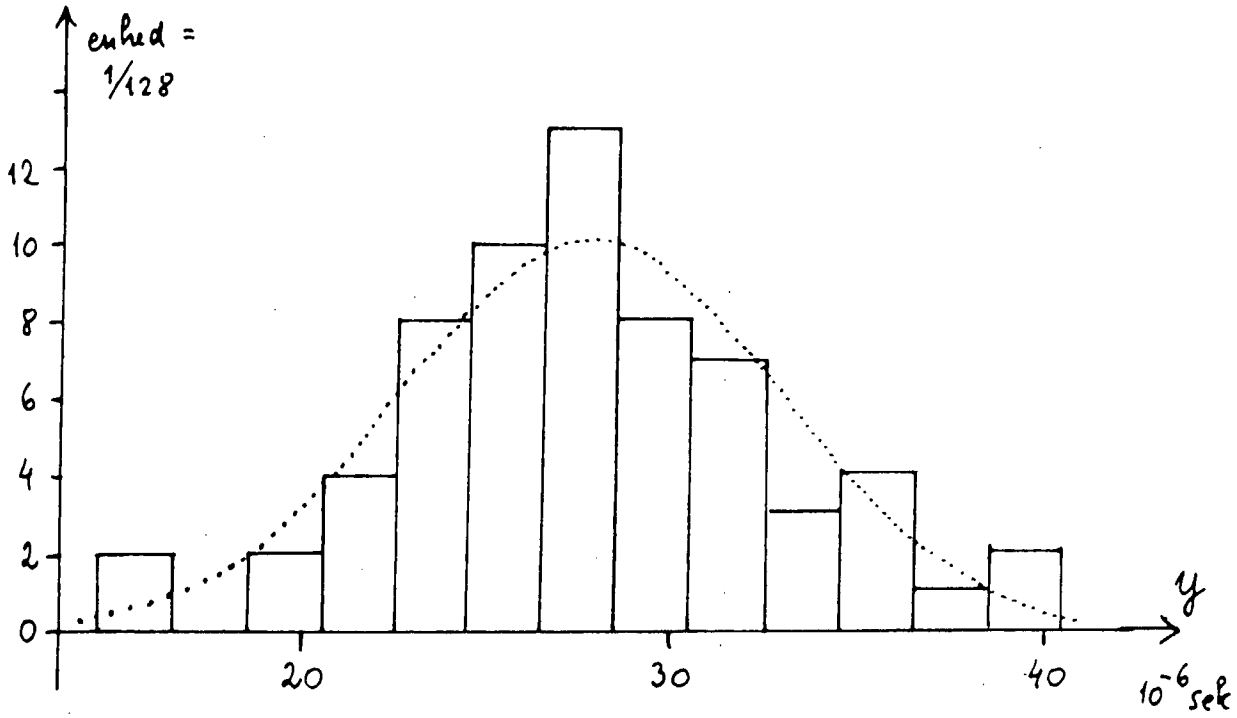
Såvel histogrammet som fraktildiagrammet viser, at det ikke er ganske urimeligt at antage at målingerne af lysets passagetid er normalfordelt.

[12] f.eks. AGF nr. 2110

Tabel 10.2: De ordnede observationer svarende til Tabel 10.1 (ekskl. -44 og -2).

y	antal	
15	0	2
16	2	
17	0	0
18	0	
19	1	2
20	1	
21	2	4
22	2	
23	3	8
24	5	
25	5	10
26	5	
27	6	13
28	7	
29	5	8
30	3	
31	2	7
32	5	
33	2	3
34	1	
35	0	4
36	4	
37	1	1
38	0	
39	1	2
40	1	
i alt	64	64

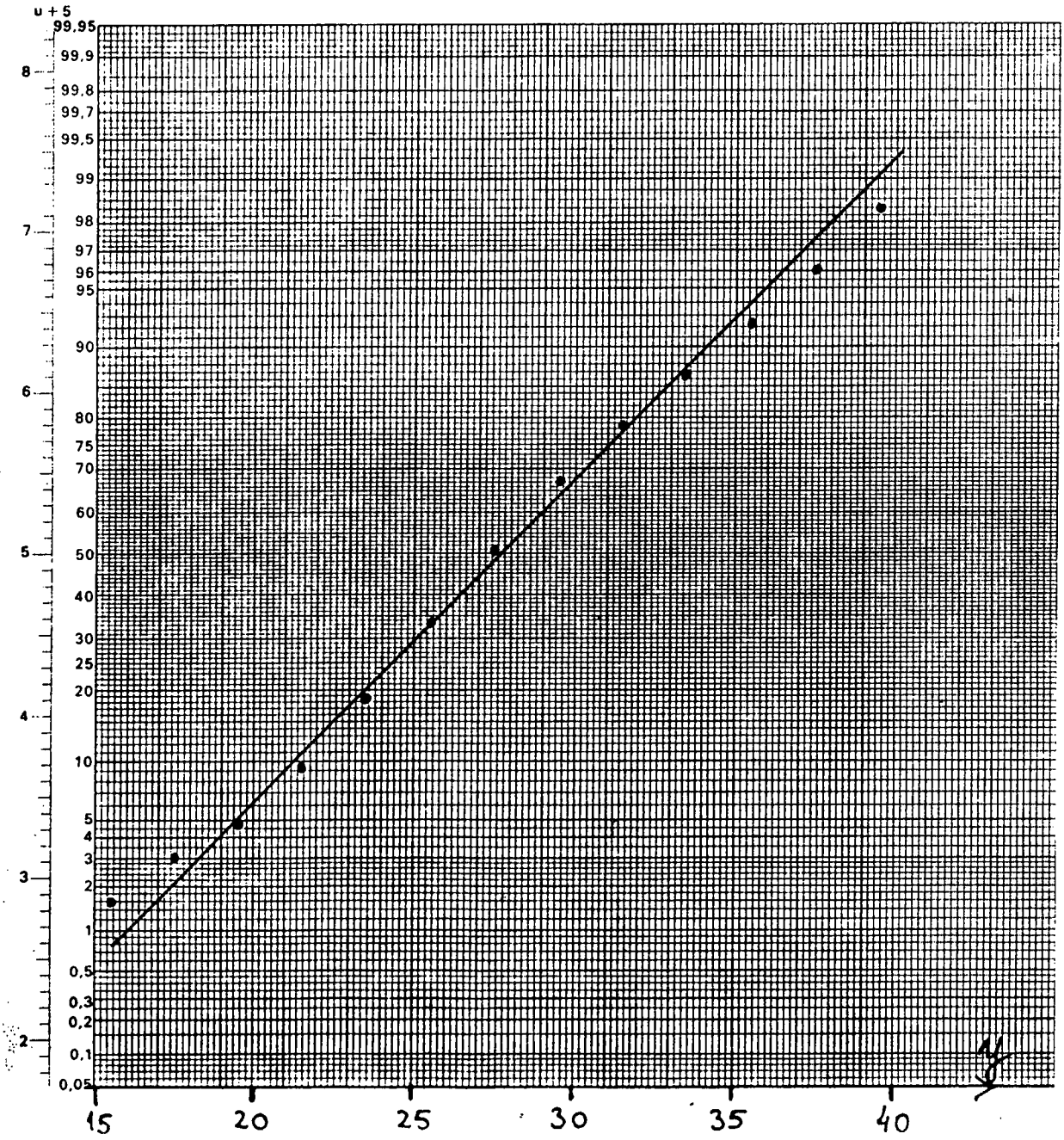
Figur 10.2: Histogram over 64 målinger af lysets passagetid. – Den prikkede kurve er tætheden for normalfordelingen med parametre $\overline{y} = 27.75$ og $s^2 = 25.8$.



Tabel 10.3: Beregninger til brug ved tegning af fraktildiagram for de grupperede lys-data.

interval-midtpunkt y	$F(y)$		
15.5	(0+ 2/2)/64	=	1.6%
17.5	(2+ 0/2)/64	=	3.1%
19.5	(2+ 2/2)/64	=	4.7%
21.5	(4+ 4/2)/64	=	9.4%
23.5	(8+ 8/2)/64	=	18.8%
25.5	(16+10/2)/64	=	32.8%
27.5	(26+13/2)/64	=	50.8%
29.5	(39+ 8/2)/64	=	67.2%
31.5	(47+ 7/2)/64	=	78.9%
33.5	(54+ 3/2)/64	=	86.7%
35.5	(57+ 4/2)/64	=	92.2%
37.5	(61+ 1/2)/64	=	96.1%
39.5	(62+ 2/2)/64	=	98.4%

Figur 10.3: Fraktildiagram over 64 målinger af lysets hastighed. Fraktildiagrammet er tegnet på grundlag af de grupperede observationer. Den indtegnede linie svarer til $\mathcal{N}(27.75\,,\,25.8)$-fordelingen.



Resumé 8. Enstikprøveproblemet i normalfordelingen

Situation: Der foreligger observationer $y_1, y_2, \ldots, y_n$ på en kontinuert måleskala.

Model: Det antages at $y_1, y_2, \ldots, y_n$ er uafhængige observationer fra normalfordelingen $\mathcal{N}(\mu, \sigma^2)$, hvor μ og σ^2 er ukendte parametre.

Estimation: Middelværdiparameteren μ estimeres ved gennemsnittet

$$\overline{y} = (y_1 + y_2 + \ldots + y_n)/n \, .$$

Variansparameteren σ^2 estimeres ved summen af de kvadratiske afvigelser divideret med antallet af frihedsgrader, dvs. ved

$$s^2 = \frac{1}{n-1} \sum_{j=1}^{n} (y_j - \overline{y})^2$$

med $f = n - 1$ frihedsgrader.

Modelkontrol: Man kan tegne et histogram (samt den fittede normalfordelingtæthed) og/eller et fraktildiagram (samt den rette linie svarende til den fittede normalfordeling).

Hypotese: Man ønsker at teste den statistiske hypotese $H_0 : \mu = \mu_0$, hvor μ_0 er et på forhånd givet tal.

Teststørrelse: Kvotienttestet kan omformes til t-testet, der er baseret på t-teststørrelsen

$$t = \frac{\overline{y} - \mu_0}{\sqrt{s^2/n}} \, .$$

Testsandsynlighed: Testsandsynligheden er sandsynligheden for at få en større $|t|$-værdi i t-fordelingen med $f = n - 1$ frihedsgrader, altså

$$\varepsilon = \mathrm{P}\left(|t_f| > |t_{\mathrm{obs}}|\right) \, ;$$

denne sandsynlighed bestemmes let ved brug af tabeller over fraktiler i t-fordelingen.

Kapitel 11

To- og flerstikprøveproblemer i normalfordelingen

En ofte forekommende situation er den, at der er foretaget målinger af en bestemt egenskab hos et antal individer, der på forhånd vides at tilhøre forskellige grupper. Alt afhængig af karakteren af målingerne kan man så benytte den ene eller anden eller tredje statistiske model/metode for dels at beskrive, dels at sammenligne de pågældende grupper. I Kapitel 4 er vist hvordan man kan sammenligne grupperne hvis det der er registreret hos de enkelte individer er et 01-svar, således at man får en binomialfordelt størrelse for hver gruppe, i Kapitel 5 behandles en situation hvor der et endeligt antal svarmuligheder for hvert enkelt individ, således at der fremkommer en multinomialfordelt størrelse for hver gruppe, og i Kapitel 8 behandles en situation hvor der for hvert individ registreres et tælletal som tænkes at stamme fra en Poissonfordeling.

I dette kapitel skal vi diskutere metoder der kan benyttes

- når der hos hvert individ er målt en enkelt talværdi,
- og når talværdien opfattes som værende en værdi på en kontinuert måleskala,
- og når man vælger at beskrive den tilfældige variation med en normalfordeling.

Årsagen til at betingelserne er formuleret med de bevidst uforpligtende verbalkonstruktioner "opfattes som værende" og "vælger at beskrive" og ikke slet og ret "er" er, at normalfordelingen meget ofte benyttes også i situationer hvor man ud fra et strengt formelt synspunkt let ville kunne pege på andre mere rigtige fordelinger. Men sagen er, at der er tit er en eller to forholdsvis gode grunde til alligevel at benytte normalfordelingen. Den ene grund er, at en sætning (Den Centrale Grænseværdisætning) fra sandsynlighedsregningen fortæller os, at *summer* af et større antal stokastiske variable under visse milde omstændigheder med god tilnærmelse er normalfordelt, og de størrelser man laver statistiske modeller for er netop tit sådanne summer. Den anden grund er rent pragmatisk: Normalfordelingsmodeller er set fra et matematisk-statistisk synspunkt meget "pæne", forstået på den måde, at når man i normalfordelingsmodeller benytter de generelle statistiske principper, så bliver resultatet næsten altid pæne og simple metoder, der ofte er lette at forstå og giver nemme og forståelige udregninger, osv. Som følge heraf er normalfordelingsmodeller studeret og beskrevet i alle ender og kanter, så man kan for det meste finde en teoretisk gennemregnet model der passer til ens behov.

Hvori består problemet?

Situationen er den, at man på hvert af et antal "individer" har målt værdien af en bestemt variabel Y. Individer skal her forstås i meget bred forstand; det kan bl.a. være personer, forsøgsdyr, jordlodder eller f.eks. de enkelte realisationer af forsøget 'at måle lysets hastighed'. Individerne er delt ind i grupper ud fra nogle kriterier som er kendt på forhånd (inden forsøget starter) og som ikke afhænger af, hvilken værdi Y nu måtte have. I den statistiske model for Y-erne vil man så regne med, at den forskel der er mellem (Y-værdierne hos) individerne *inden for* en bestemt gruppe er *tilfældig*, og at den forskel der er *mellem* forskellige grupper er *systematisk*. En normalfordelingsmodel til denne situation er da indrettet på den måde, at

- *den systematiske forskel* mellem grupper beskrives ved hjælp af middelværdiparametre, og
- *den tilfældige forskel* inden for grupper beskrives ved hjælp af dels normalfordelingen, dels variansparametre i normalfordelingen.

Det statistiske problem består tit i, at man ønsker at sammenligne grupperne for at vurdere om den systematiske forskel mellem grupperne er signifikant, dvs. om den forskel der er mellem grupperne er stor målt i forhold til den tilfældige variation der er inden for de enkelte grupper. Man ønsker derfor i en vis forstand at måle forskellen mellem grupperne med en målestok der er kalibreret efter størrelsen af den tilfældige variation inden for grupperne.

Det man egentlig er interesseret i er altså information om middelværdiparametrene. Men for at der kan være en veldefineret målestok at måle dem med, må man først sikre sig at det har mening at tale om *den* tilfældige variation inden for grupper. Derfor man i må modellen gøre den antagelse (som undertiden kan testes), at *de forskellige grupper har samme variansparameter*[1].

Hermed er problemet beskrevet. I resten af kapitlet skal vi se hvordan det løses. Men først præsenteres et eksempel.

Eksempel 11.1. ***Dækningsgrad for Fuglegræs***

På dyrkede marker er ukrudt jo pr. definition en uting, og landmanden kan overveje om han skal sprøjte mod den slags ukrudt han anser for værst. Men når man fjerner én slags ukrudt, kan det jo være, at det ikke bare er afgrøden der derved får forbedrede vækstforhold, men også de resterende ukrudtsarter! Måske er det en fordel at have så mange forskellige ukrudtsarter som muligt, fordi de så kan holde hinanden i skak. For at undersøge ukrudtsplanters indbyrdes konkurrence på en kornmark har man[2] *udført et større forsøg, der går ud på, at på forskellige dele af en stor mark luger man på et bestemt tidspunkt forskellige ukrudtsarter bort, og derefter ser man, hvorledes resten af arterne så trives. Mere præcist er marken delt op i 16 jordlodder, som er delt ind i fire grupper med hver fire lodder. Den første gruppe er en kontrolgruppe hvor intet luges bort, men i hver af grupperne to, tre og fire luges én bestemt ukrudtsart bort (henholdsvis Snerle pileurt, Fuglegræs og Hvidmelet gåsefod). Een gang før og tre gange efter bortlugningen registrerer man hvilke planter der*

[1] Man kan dog klare sig med antagelsen om, at gruppernes variansparametre er kendte pånær en konstant faktor.

[2] A. Greenfort, C.S.F. Jensen & S. Jeppesen: *Planter og planter imellem*. RUC, 1987.

Tabel 11.1: Dækningsgrader for Fuglegræs ved første registrering.

gruppe	dækningsgrader			
1	17	38	23	26
2	19	16	16	14
3	25	33	29	33
4	27	16	30	20

er på de forskellige lodder og i hvor stor udstrækning. Den første registrering skal tjene til at fastlægge det niveau hvorudfra den senere udvikling skal måles.

De fire grupper er fordelt på marken i et romersk kvadrat:

3	*4*	*1*	*2*
2	*1*	*4*	*3*
4	*3*	*2*	*1*
1	*2*	*3*	*4*

De fire lodder der udgør en gruppe er altså placeret fire helt forskellige steder på marken; derved har man en chance for at kunne tage højde for eventuelle variationer i jordbund og mikroklima henover marken.

Forsøget har givet et stort talmateriale, som kan analyseres på mange måder. Her skal vi kun se på en enkelt detalje i forbindelse med fastlæggelsen af et udgangsniveau på grundlag af den første registrering. Vi vil studere forekomsten af Fuglegræs (Stellaria media) *ved den første registrering, se Tabel 11.1. Registreringen foregår ved hjælp af et gitternet med 416 gitterpunkter placeret som på et almindeligt stykke ternet papir, dog er sidelængden i ternerne 5 cm. Registreringen foregår på den måde, at gitternettet placeres på jordlodden, hvorefter man i hvert gitterpunkt ser efter, om der findes noget af en Fuglegræs-plante eller ej. Som mål for* dækningsgraden *for arten benyttes* antallet af gitterpunkter hvor arten blev registreret. *Dækningsgraden bliver på denne måde et helt tal mellem 0 og 416.*

Da den første registrering udførtes inden *der blev foretaget nogen bortlugning, kan der ikke på dette tidspunkt være tale om nogen*

behandlingseffekt (lugningseffekt). De forskelle der er på lodderne og på grupperne må alene skyldes "startbetingelserne", dvs. de lokale variationer i jordbund og klima og de forskellige antal planter af den pågældende art som der nu tilfældigvis var på de enkelte områder af marken. Da man ønsker at vurdere hvordan behandlingerne påvirker grupperne, kan det være af interesse at få en idé om, hvor forskellige (eller hvor ens) grupperne egentlig er ved forsøgets start. Hvis grupperne nemlig er rimeligt ens, kan man bestemme et fælles startniveau hvorudfra den senere udvikling kan vurderes, men hvis der er en signifikant forskel mellem grupperne, så er man nødt til at vurdere hver gruppes udvikling ud fra dens eget startniveau. Derfor vil vi gerne sammenligne de fire grupper og vurdere, om forskellen mellem grupperne er stor i forhold til den tilfældige variation inden for grupperne.

Den statistiske model:
Da observationerne er fremkommet som en sum af et vist antal 01-størrelser (svarende til om planten er fraværende eller til stede i det pågældende gitterpunkt), kunne man sige at det smager lidt af en binomialfordelingssituation (eller eventuelt en Poissonfordelingssituation da n er temmelig stor). Hertil må man sige, at ikke alle binomialfordelingsbetingelserne er opfyldt, idet de enkelte 01-størrelser næppe er uafhængige med samme sandsynlighed for '1'. Der vil derfor formentlig være en større tilfældig variation inden for de enkelte grupper end hvad binomialfordelingen kan forklare. Man kan derfor, idet man går let hen over at der er tale om diskrete observationer, forsøge sig med en normalfordelingmodel, hvor man jo ved hjælp af variansparameteren kan modellere den tilfældige variation særskilt. Vi vil derfor benytte en statistisk model der går ud på, at observationer i samme gruppe opfattes som observationer fra en og samme normalfordeling, og at de fire grupper har hver deres normalfordeling. Det statistiske problem er da at undersøge, om de fire normalfordelinger kan tænkes at være ens. □

11.1 k-stikprøveproblemet i normalfordelingen

Antag at der foreligger nogle observationer y som er ordnet i k grupper med n_i observationer i gruppe nr. i, $i = 1, 2, \ldots, k$. Observation nr. j fra gruppe nr. i betegnes y_{ij}. Skematisk ser det sådan ud:

gruppe	observationer					
1	y_{11}	y_{12}	$\cdots$	y_{1j}	$\cdots$	y_{1n_1}
2	y_{21}	y_{22}	$\cdots$	y_{2j}	$\cdots$	y_{2n_2}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
i	y_{i1}	y_{i2}	$\cdots$	y_{ij}	$\cdots$	y_{in_i}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
k	y_{k1}	y_{k2}	$\cdots$	y_{kj}	$\cdots$	y_{kn_k}

Vi går ud fra, at forskellen mellem observationerne inden for en gruppe er tilfældig, hvorimod der er en systematisk forskel mellem grupperne. Vi går endvidere ud fra, at y_{ij}-erne er observerede værdier af uafhængige stokastiske variable Y_{ij}. Den tilfældige variation vil vi beskrive ved hjælp af en normalfordeling, og det skal derfor alt i alt være sådan at Y_{ij} *er normalfordelt med middelværdi* μ_i *og varians* σ^2, kort

$$Y_{ij} \sim \mathcal{N}(\mu_i, \sigma^2) . \tag{11.1}$$

Herved beskriver middelværdiparametrene $\mu_1, \mu_2, \ldots, \mu_k$ *den systematiske variation*, nemlig de enkelte gruppers niveauer, medens variansparameteren σ^2 (samt normalfordelingen) beskriver *den tilfældige variation* inden for grupperne; denne tilfældige variation antages at være den samme i alle grupperne, jf. den tidligere diskussion (side 201).

Estimation af parametrene

De ukendte middelværdiparametre $\mu_1, \mu_2, \ldots, \mu_k$ i grundmodellen (11.1) estimeres ved hjælp af maximum likelihood metoden, altså som de værdier der maksimaliserer likelihoodfunktionen, der er

$$L(\mu_1, \mu_2, \ldots, \mu_k, \sigma^2) = \prod_{i=1}^{k} \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\tfrac{1}{2}\frac{(y_{ij} - \mu_i)^2}{\sigma^2}\right)$$

$$= (2\pi)^{-n/2}(\sigma^2)^{-n/2}\exp\left(-\frac{1}{2}\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\mu_i)^2}{\sigma^2}\right), \qquad (11.2)$$

hvor $n = n_1 + n_2 + \ldots + n_k$ er det samlede antal observationer. Af udtrykket (11.2) fremgår, at hvis σ^2 er fast så er opgaven at maksimalisere likelihoodfunktionen L med hensyn til $\mu_1, \mu_2, \ldots, \mu_k$ den samme som opgaven at minimalisere kvadratsummen

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\mu_i)^2 ,$$

og den opgave er ret let at løse:

Vi lader $\overline{y}_i$ betegne gennemsnittet i gruppe i,

$$\overline{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} .$$

Vi benytter så formlen for kvadratet på en toleddet størrelse:

$$\begin{aligned}(y_{ij}-\mu_i)^2 &= \big((y_{ij}-\overline{y}_i)+(\overline{y}_i-\mu_i)\big)^2 \\ &= (y_{ij}-\overline{y}_i)^2 + 2(y_{ij}-\overline{y}_i)(\overline{y}_i-\mu_i) + (\overline{y}_i-\mu_i)^2 .\end{aligned}$$

Når vi her holder i fast og summerer over j bliver summen af de dobbelte produkter 0, fordi $\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_i)$ er lig med 0 ifølge definitionen af $\overline{y}_i$. Hvis vi endelig også summerer over i får vi

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\mu_i)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_i)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{y}_i-\mu_i)^2 .$$

Opgaven er at minimalisere venstresiden. Men de μ-er der minimaliserer venstresiden er de samme som dem der minimaliserer den anden kvadratsum på højresiden, og den bliver mindst mulig, nemlig 0, netop når μ_i er lig $\overline{y}_i$, $i = 1, 2, \ldots, k$. Vi har dermed fundet at maksimaliseringsestimatet for den i-te gruppes middelværdi er lig med gennemsnittet af observationerne i gruppen, $\widehat{\mu}_i = \overline{y}_i$.

Derefter går vi over til at estimere σ^2. Maksimaliseringsestimatet $\widehat{\sigma^2}$ for σ^2 kan bestemmes som maksimumspunktet for funktionen

$$\sigma^2 \;\mapsto\; L(\overline{y}_1, \overline{y}_2, \ldots, \overline{y}_k, \sigma^2) ;$$

man finder da at

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2 .$$

En størrelse som $y_{ij} - \overline{y}_i$ der er forskellen mellem den faktiske observation og det bedst mulige 'fit' under den aktuelle model, kaldes undertiden for et *residual*[3]. Derfor kaldes en størrelse som

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2$$

for en *residualkvadratsum*, og man kan sige at maksimaliseringsestimatet $\widehat{\sigma^2}$ for σ^2 er lig med residualkvadratsummen divideret med antallet af observationer. Som regel benytter man imidlertid et andet skøn over σ^2, nemlig residualkvadratsummen divideret med *antallet af frihedsgrader* $n - k$ (antal af observationer minus antal af estimerede parametre), dvs. man benytter variansskønnet

$$s_0^2 = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2 .$$

Man begrunder brugen af s_0^2 frem for $\widehat{\sigma^2}$ på lignende måde som i Enstikprøveproblemet i normalfordelingen, se side 186.

Sammenfattende har vi altså at

- middelværdiparameteren μ_i i den i-te gruppe estimeres ved gennemsnittet $\overline{y}_i$ af observationerne i gruppen,
- den fælles varians σ^2 for grupperne estimeres ved residualkvadratsummen divideret med antallet af frihedsgrader,

$$s_0^2 = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2 \qquad (11.3)$$

  med $n - k$ frihedsgrader.

I Tabel 11.2 er vist de værdier man finder i Fuglegræs-eksemplet.

[3]Et *residual* betyder: noget der er til rest.

Tabel 11.2: Fuglegræs-eksemplet: nogle beregnede størrelser.

i	n_i	$y_{i\bullet} = \sum_{j=1}^{n_i} y_{ij}$	$\overline{y}_i$	$\sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)^2$
1	4	104	26.00	234.00
2	4	65	16.25	12.75
3	4	120	30.00	44.00
4	4	93	23.25	122.75
sum	16	382		413.50
gennemsnit			23.88	

$$s_0^2 = \frac{1}{16-4} 413.50 = 34.46 .$$

11.2 Bartlett's test for varianshomogenitet

Som nævnt[4] er det i normalfordelingsmodeller en forudsætning for en meningsfuld sammenligning af middelværdiparametre for forskellige grupper, at disse grupper har samme varians[5]. I dette afsnit skal vi omtale et test for, om et antal grupper af normalfordelte observationer kan antages at have samme varians, dvs. om der er *varianshomogenitet*. Testet kan ikke benyttes hvis nogen af grupperne kun indeholder én observation, og for at man skal kunne anvende den tilnærmede fordeling af teststørrelsen skal hver gruppe indeholde mindst seks observationer[6].

Antag at situationen er som beskrevet i Afsnit 11.1 og at vi ønsker at teste antagelsen om at grupperne har samme variansparameter σ^2. Den måde man kan teste en sådan antagelse på er, at man indlejrer den statistiske model i en større model, og så tester man, på helt sædvanlig måde, om man kan reducere den store model til den oprindelige model.

[4] på side 201

[5] Man kan eventuelt klare sig med en antagelse om, at gruppernes varianser er af formen: en ukendt fælles parameter ganget med en kendt konstant (der kan afhænge af gruppen).

[6] eller rettere: i hver enkelt gruppe skal variansskønnet have mindst fem frihedsgrader.

I det aktuelle tilfælde indlejrer vi den oprindelige model (11.1) fra side 204 i den større model der tillader grupperne at have hver deres egen varians, nemlig

$$Y_{ij} \sim \mathcal{N}(\mu_i, \sigma_i^2) . \tag{11.4}$$

Dernæst tester vi (11.1) på side 204 som en hypotese i forhold til den nye grundmodel (11.4).

Den hypotese der skal testes handler kun om en del af modellens parametre, og for så at sige at slippe af med de parametre der ikke har noget med hypotesen at gøre (altså med μ_i-erne) plejer man at teste hypotesen i den betingede fordeling givet skønnene over middelværdiparametrene[7]. Hvis man omskriver kvotienttesteststørrelsen i den nævnte betingede fordeling når man frem til at følgende størrelse (*Bartlett's teststørrelse*) kan benyttes som teststørrelse for hypotesen om varianshomogenitet:

$$B = -\sum_{i=1}^{k} f_i \ln \frac{s_i^2}{s_0^2} ; \tag{11.5}$$

her betegner s_i^2 skønnet over variansen σ_i^2 i den i-te gruppe og f_i antallet af frihedsgrader for s_i^2, dvs.

$$s_i^2 = \frac{1}{f_i} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)^2,$$

$$f_i = n_i - 1 ,$$

og s_0^2 er det gamle skøn (11.3) over den fælles varians σ^2. Bemærk at s_0^2 er et vægtet gennemsnit af s_i^2-erne med frihedsgradsantallene som vægte,

$$s^2 = \frac{1}{f} \sum_{i=1}^{k} f_i s_i^2$$

hvor $f = f_1 + f_2 + \ldots + f_k$ er antallet af frihedsgrader for s_0^2.

Teststørrelsen B (som i virkeligheden er en $-2 \ln Q$-størrelse) er altid et positivt tal, og store værdier af B er signifikante, dvs. tyder på at hypotesen om varianshomogenitet er forkert. Hvis hypotesen er

[7]Dette hænger sammen med at man også *estimerer* variansparametrene i den betingede fordeling givet middelværdiskønnene, jf. side 186.

Tabel 11.3: Fuglegræs-eksemplet: nogle beregnede størrelser.

n står for antal observationer y, S for Sum af y-er, $\overline{y}$ for gennemsnit af y-er, f for antal frihedsgrader, SS for Sum af kvadratiske afvigelser ('Sum of Squared deviations'), og s^2 for variansskøn (SS/f).

gruppe	n	S	$\overline{y}$	f	SS	s^2
1	4	104	26.00	3	234.00	78.00
2	4	65	16.25	3	12.75	4.25
3	4	120	30.00	3	44.00	14.67
4	4	93	23.25	3	122.75	40.92
sum	16	382		12	413.50	
gennemsnit			23.88			34.46

rigtig, er B nogenlunde χ^2-fordelt med $k-1$ frihedsgrader, således at det er let at bestemme den omtrentlige testsandsynlighed som

$$\varepsilon \;=\; \mathrm{P}\left(\chi^2_{k-1} \geq B_{\mathrm{obs}}\right) .$$

Denne χ^2-approksimation er god når alle f_i-erne er store; som tommelfingerregel siger man, at de alle skal være mindst 5.

Eksempel 11.2. ***Fuglegræs: test for varianshomogenitet***

Som illustration udregnes Bartlett's teststørrelse i Fuglegræseksemplet. Vi udvider det tidligere regneskema i Tabel 11.2 og får Tabel 11.3. Derefter kan vi udregne B_{obs}*:*

$$\begin{aligned} B_{\mathrm{obs}} &= -\left(3\,\ln\frac{78.00}{34.46} + 3\,\ln\frac{4.25}{34.46} + 3\,\ln\frac{14.67}{34.46} + 3\,\ln\frac{40.92}{34.46}\right) \\ &= 5.87\,. \end{aligned}$$

Betingelsen om at alle f_i*-erne skal være mindst fem er ikke opfyldt (idet de alle er tre), så det er begrænset hvor* χ^2*-fordelt* B *kan forventes at være; men hvis vi ser lidt stort på det, så skulle* B *altså være ca.* χ^2*-fordelt med* $k-1=4-1=3$ *frihedsgrader når hypotesen om varianshomogenitet er rigtig. I* χ^2_3*-fordelingen er 80%-fraktilen 4.64 og 90%-fraktilen 6.25, således at under forudsætning af at hypotesen er rigtig er der i størrelsesordenen 10%*

sandsynlighed for at få en værre B-værdi end den opnåede; på dette grundlag vil vi ikke forkaste hypotesen om varianshomogenitet. □

Tilfældet $k = 2$

Hvis der kun er to grupper kan man teste hypotesen om varianshomogenitet på en simplere måde, idet man så ganske enkelt kan benytte teststørrelsen

$$R = \frac{s_1^2}{s_2^2},$$

altså forholdet mellem de to variansskøn. R-størrelsen er et positivt tal, og værdier tæt på 1 tyder på at hypotesen om varianshomogenitet er god nok, hvorimod såvel meget store som meget små R-værdier er signifikante. Der er tale om et *tosidet test*, hvor man som testsandsynlighed ε benytter sandsynligheden (når hypotesen er rigtig) for at få en R-værdi der ligger uden for intervallet med endepunkter R_{obs} og $1/R_{\text{obs}}$, dvs.

$$\begin{aligned} \varepsilon &= \mathrm{P}_0\left(R > R_{\text{obs}}\right) + \mathrm{P}_0\left(R < \frac{1}{R_{\text{obs}}}\right) \\ &= \mathrm{P}_0\left(R > R_{\text{obs}}\right) + \mathrm{P}_0\left(\frac{1}{R} > R_{\text{obs}}\right) \end{aligned} \tag{11.6}$$

hvis $R_{\text{obs}} > 1$, og

$$\begin{aligned} \varepsilon &= \mathrm{P}_0\left(R < R_{\text{obs}}\right) + \mathrm{P}_0\left(R > \frac{1}{R_{\text{obs}}}\right) \\ &= \mathrm{P}_0\left(\frac{1}{R} > \frac{1}{R_{\text{obs}}}\right) + \mathrm{P}_0\left(R > \frac{1}{R_{\text{obs}}}\right) \end{aligned} \tag{11.7}$$

hvis $R_{\text{obs}} < 1$.

Man kan vise, at når hypotesen om varianshomogenitet er rigtig, så vil R-teststørrelsen følge en F-fordeling med (f_1, f_2) frihedsgrader. Da man har tabeller over fraktiler i F-fordelingen[8] er det derfor let at bestemme testsandsynligheden ε. Hvis man yderligere udnytter en særlig

[8] F-fordelingen hedder i visse tabelværker v^2-fordelingen.

egenskab ved F-fordelinger, nemlig at hvis R følger F_{f_1,f_2}-fordelingen så vil $1/R$ følge F_{f_2,f_1}-fordelingen, så kan de to udtryk (11.6) og (11.7) forsimples til

$$\varepsilon = \mathrm{P}\left(F_{f_1,f_2} > R_{\mathrm{obs}}\right) + \mathrm{P}\left(F_{f_2,f_1} > R_{\mathrm{obs}}\right)$$

når $R_{\mathrm{obs}} > 1$, og

$$\varepsilon = \mathrm{P}\left(F_{f_2,f_1} > \frac{1}{R_{\mathrm{obs}}}\right) + \mathrm{P}\left(F_{f_1,f_2} > \frac{1}{R_{\mathrm{obs}}}\right)$$

når $R_{\mathrm{obs}} < 1$. Det er disse udtryk man anvender i praksis.

Resumé 9. Bartlett's test for varianshomogenitet

Situation: Der foreligger k grupper af normalfordelte observationer, således at observationer i samme gruppe stammer fra samme normalfordeling. I hver gruppe er udregnet et variansskøn; variansskønnet fra gruppe i betegnes s_i^2 og har f_i frihedsgrader.

Hypotese: Man ønsker at teste den statistiske hypotese om at de k grupper har samme varians.

Estimation: Hvis hypotesen er rigtig, så skal den fælles varians estimeres som et vægtet gennemsnit af de enkelte gruppers varianser, nemlig ved

$$s_0^2 = \frac{1}{f}\sum_{i=1}^{k} f_i s_i^2$$

der har $f = f_1 + f_2 + \ldots + f_k$ frihedsgrader.

Teststørrelse: Som teststørrelse anvendes

$$B = -\sum_{i=1}^{k} f_i \ln \frac{s_i^2}{s_0^2} .$$

Store værdier af B er signifikante.

Testsandsynlighed: Når $f_1, f_2, \ldots, f_k$ alle er mindst 5, så er B med god tilnærmelse χ^2-fordelt med $k-1$ frihedsgrader når hypotesen er rigtig, og testsandsynligheden kan da findes som

$$\varepsilon = \mathrm{P}\left(\chi_{k-1}^2 \geq B_{\mathrm{obs}}\right) .$$

Specialtilfældet $k = 2$: Når der kun er to variansskøn s_1^2 og s_2^2 der skal sammenlignes kan det ske ved at udregne kvotienten

$$R = \frac{s_1^2}{s_2^2} ,$$

der er F-fordelt med frihedsgradsantal f_1 og f_2 når hypotesen er rigtig. R-værdier langt fra 1 er signifikante, og man plejer at bestemme testsandsynligheden ε som sandsynligheden for at få en F_{f_1,f_2}-værdi der ligger uden for intervallet med endepunkter R_{obs} og $1/R_{\mathrm{obs}}$.

11.3 Ensidet variansanalyse

Det statistiske hovedproblem i dette kapitel er sammenligning af k grupper af normalfordelte observationer, idet vi benytter (11.1) på side 204 som grundmodel. I dette afsnit skal vi beskæftige os med spørgsmålet om, hvordan man undersøger om de k grupper kan antages at have samme middelværdi. Opgaven er således at teste hypotesen H_0 om, at der ikke er nogen signifikant forskel mellem grupperne, også kaldet hypotesen om *homogenitet mellem grupper*:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k \ .$$

Meget ofte er det ikke H_0 man er interesseret i, men dens modsætning at der *er* en signifikant forskel mellem grupperne, fordi man netop har et ønske eller et håb om at kunne vise at grupperne *ikke* er ens. Når det alligevel er H_0 man tester og ikke dens modsætning, så hænger det sammen med to generelle træk ved formulering og test af statistiske hypoteser:

1. De hypoteser man kan teste er altid hypoteser der går ud på en *forsimpling* af den aktuelle grundmodel – typisk tester man at nogle parametre er ens i forhold til en grundmodel, der tillader dem at være forskellige.

2. Det er informativt at få *forkastet* en hypotese: Vi får at vide at der er en signifikant uoverensstemmelse mellem hypotese og observationer.

   Derimod viser det ofte ingenting at få accepteret en hypotese: Det kan være at man simpelt hen bare har for få observationer til at kunne afsløre noget som helst.

Vi skal nu se hvordan man tester hypotesen H_0 om ens middelværdier. Man kan gå frem efter den sædvanlige opskrift, dvs. opstille en kvotientteststørrelse der sammenligner likelihoodfunktionens maksimale værdier hhv. under H_0 og under grundmodellen. Vi har tidligere (side 205) fundet, at likelihoodfunktionen (11.2) maksimaliseres af værdierne $\overline{y}_1, \overline{y}_2, \ldots, \overline{y}_k, \widehat{\sigma^2}$ hvor $\widehat{\sigma^2} = \frac{1}{n}\sum\sum(y_{ij} - \overline{y}_i)^2$. Nu skal vi bestemme de værdier der maksimaliserer likelihoodfunktionen under H_0. Under H_0 er der ingen forskel på grupperne, men det hele er i realiteten én stor gruppe med n observationer, og fra behandlingen af enstikprøveproblemet i normalfordelingen vides (side 184f) at maksimaliseringsestimatet

for den fælles middelværdi μ er det totale gennemsnit $\overline{y}$ og for den fælles varians σ^2 er det kvadratafvigelsessummen omkring $\overline{y}$ divideret med n, dvs. $\widehat{\widehat{\sigma^2}} = \frac{1}{n}\sum\sum(y_{ij} - \overline{y})^2$.

Kvotienttesteststørrelsen for H_0 er derfor

$$\begin{aligned} Q &= \frac{L(\overline{y}, \overline{y}, \ldots, \overline{y}, \widehat{\widehat{\sigma^2}})}{L(\overline{y}_1, \overline{y}_2, \ldots, \overline{y}_k, \widehat{\sigma^2})} \\ &= \left(\frac{\widehat{\widehat{\sigma^2}}}{\widehat{\sigma^2}}\right)^{-n/2} \exp\left(-\frac{1}{2}\left(\frac{\sum\sum(y_{ij} - \overline{y})^2}{\widehat{\widehat{\sigma^2}}} - \frac{\sum\sum(y_{ij} - \overline{y}_i)^2}{\widehat{\sigma^2}}\right)\right) \\ &= \left(\frac{\sum\sum(y_{ij} - \overline{y})^2}{\sum\sum(y_{ij} - \overline{y}_i)^2}\right)^{-n/2} . \end{aligned}$$

For at kunne omskrive Q yderligere skal vi bruge en opspaltning af kvadratsummen i tælleren. For det første er

$$\begin{aligned} (y_{ij} - \overline{y})^2 &= \big((y_{ij} - \overline{y}_i) + (\overline{y}_i - \overline{y})\big)^2 \\ &= (y_{ij} - \overline{y}_i)^2 + 2(y_{ij} - \overline{y}_i)(\overline{y}_i - \overline{y}) + (\overline{y}_i - \overline{y})^2 . \end{aligned}$$

Når vi her holder i fast og summerer over j, bliver summen af de dobbelte produkter 0 fordi $\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i) = 0$; alt i alt får vi dermed denne opspaltning af kvadratsummen i tælleren:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{y}_i - \overline{y})^2 , \quad (11.8)$$

dvs. den totale kvadratsum, der beskriver y_{ij}-ernes variation om det totale gennemsnit $\overline{y}$, spaltes op i en sum af dels et bidrag der beskriver "variationen inden for grupperne", dels et bidrag der beskriver "variationen mellem grupperne".

Parallelt med spaltningen af kvadratsummen har vi opspaltningen

$$n - 1 = (n - k) + (k - 1)$$

af frihedsgraderne, og ved at dividere kvadratsummerne med de tilsvarende antal frihedgrader får vi variansskøn der beskriver forskellige variationer:

- *Variationen omkring totalgennemsnittet* (dvs. enkeltobservationernes variation omkring totalgennemsnittet) beskrives af variansskønnet under H_0

$$s_{01}^2 = \frac{1}{n-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\overline{y})^2 .$$

- *Variationen inden for grupper* (dvs. enkeltobservationernes variation omkring deres respektive gruppegennemsnit) beskrives af variansskønnet (11.3) i grundmodellen (11.1)

$$s_0^2 = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_i)^2 .$$

- *Variationen mellem grupper* (dvs. gruppegennemsnittenes variation omkring det totale gennemsnit) beskrives af

$$\begin{aligned} s_1^2 &= \frac{1}{k-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{y}_i-\overline{y})^2 \\ &= \frac{1}{k-1}\sum_{i=1}^{k} n_i(\overline{y}_i-\overline{y})^2 . \end{aligned}$$

Men vi skal videre med omskrivningen af udtrykket for Q. Ved hjælp af (11.8) kan vi omskrive det til

$$\begin{aligned} Q &= \left(1+\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{y}_i-\overline{y})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_i)^2}\right)^{-n/2} \\ &= \left(1+\frac{(k-1)s_1^2}{(n-k)s_0^2}\right)^{-n/2} , \end{aligned}$$

og af det kan man se at Q er en monotont aftagende funktion af størrelsen

$$F = \frac{s_1^2}{s_0^2} , \qquad (11.9)$$

således at store værdier af F svarer til små værdier af Q og er dermed tegn på at H_0 bør forkastes.

Som teststørrelse for H_0 benytter man i praksis altid F. Det ses at F kan forstås som *forholdet mellem variationen mellem grupper og variationen inden for grupper.*

Man forkaster hypotesen om homogenitet mellem grupper når variationen *mellem* grupper er væsentlig større end variationen *inden for* grupper. Man kan bevise, at forudsat at hypotesen H_0 er rigtig er F-teststørrelsen F-fordelt med frihedsgrader $(k-1, n-k)$, hvilket betyder at testsandsynligheden

$$\varepsilon \;=\; \mathrm{P}_0\left(F > F_{\mathrm{obs}}\right)$$

kan bestemmes som

$$\varepsilon \;=\; \mathrm{P}\left(F_{k-1,n-k} > F_{\mathrm{obs}}\right)$$

der let findes ved hjælp af en tabel over fraktiler i F-fordelingen[9].

Vi har hermed løst den opgave der gik ud på at sammenligne k grupper af normalfordelte observationer. Det ses at selve testet (11.9) går ud på at sammenligne to variansskøn, og derfor kan man sige at vi foretager en *variansanalyse*; da observationerne er er inddelt efter ét kriterium (nemlig hvilken gruppe de tilhører), kaldes analysen for *ensidet variansanalyse*. Det er kutyme at give en oversigt over en variansanalyse i et såkaldt *variansanalyseskema*. Tabel 11.4 er et variansanalyseskema for Fuglegræs-eksemplet.

Eksempel 11.3. ***Fuglegræs, konklusion***

Tabel 11.4 viser variansanalyseskemaet for ensidet variansanalyse i fuglegræseksemplet. Det ses at F-teststørrelsen bliver 3.9, og denne værdi skal sammenholdes med fraktilerne i F-fordelingen med frihedsgrader 3 og 12; i denne fordeling er 95%-fraktilen 3.49 og 97.5%-fraktilen 4.47, så testsandsynligheden er knap 4 %. På den baggrund vil man sædvanligvis være stemt for at forkaste *hypotesen om ens middelværdier i grupperne. Man må altså konstatere, at de fire grupper synes at være forskellige allerede inden man begynder at give dem hver deres behandling. Det kan måske*

[9] I visse tabelværker kaldes F-fordelingen for v^2-fordelingen.

Tabel 11.4: Fuglegræs-eksemplet: *Variansanalyseskema.*

f står for antal frihedsgrader, SS for Sum af kvadratiske afvigelser, $s^2 = SS/f$.

variation	f	SS	s^2	test
inden for grupper	12	413.50	34.46	
mellem grupper	3	402.25	134.08	134.08/34.46=3.9
total	15	815.75	54.38	

synes overraskende, men det må hænge sammen med at der på forhånd er betydelige forskelle på de forskellige dele af marken. Når man sidenhen skal undersøge hvordan behandlingerne virker, er man nødt til at tage hensyn til denne forskellighed. □

Resumé 10. Ensidet variansanalyse

Situation: Der foreligger nogle observationer y som er målt på en kontinuert måleskala og som er inddelt i k grupper, således at der er n_i observationer i gruppe nr. i, $i = 1, 2, \ldots, k$. Observation nr. j fra gruppe nr. i betegnes y_{ij}. Skematisk ser det således ud:

gruppe	observationer					
1	y_{11}	y_{12}	$\cdots$	y_{1j}	$\cdots$	y_{1n_1}
2	y_{21}	y_{22}	$\cdots$	y_{2j}	$\cdots$	y_{2n_2}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
i	y_{i1}	y_{i2}	$\cdots$	y_{ij}	$\cdots$	y_{in_i}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
k	y_{k1}	y_{k2}	$\cdots$	y_{kj}	$\cdots$	y_{kn_k}

Model: Det antages at y_{ij}-erne er observerede værdier af uafhængige stokastiske variable Y_{ij}, således at Y_{ij} er normalfordelt med middelværdi μ_i og varians σ^2, kort

$$Y_{ij} \sim \mathcal{N}(\mu_i, \sigma^2),$$

hvor $\mu_1, \mu_2, \ldots, \mu_k$ og σ^2 er ukendte parametre. Herved beskriver middelværdiparametrene $\mu_1, \mu_2, \ldots, \mu_k$ den systematiske variation, nemlig de enkelte gruppers niveauer, medens variansparameteren σ^2 (samt normalfordelingen) beskriver den tilfældige variation inden for grupperne; den tilfældige variation antages at være den samme i alle grupper.

Estimation: Middelværdiparametrene $\mu_1, \mu_2, \ldots, \mu_k$ estimeres ved gruppegennemsnittene:

$$\begin{aligned} \hat{\mu}_i &= \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \\ &= \overline{y}_i \,. \end{aligned}$$

Variansparameteren σ^2 estimeres som den gennemsnitlige kvadratiske afvigelse mellem observationerne og deres gruppegen-

nemsnit, nemlig

$$s_0^2 = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\overline{y}_i)^2$$

der har $f = n - k$ frihedsgrader ($n = n_1 + \ldots + n_k$).

Modelkontrol: Hvis der er tilstrækkelig mange observationer i grupperne kan man for hver gruppe tegne et histogram (samt den fittede normalfordelingtæthed) og/eller et fraktildiagram (samt den rette linie svarende til den fittede normalfordeling).

Antagelsen om varianshomogenitet kan eventuelt testes med Bartlett's test for varianshomogenitet.

Hypotese: Man ønsker at teste hypotesen om homogenitet mellem grupper, dvs. den statistiske hypotese

$$H_0: \ \mu_1 = \mu_2 = \ldots = \mu_k$$

om at der ikke er nogen signifikant forskel mellem grupperne.

Teststørrelse: Udregn variationen mellem grupper:

$$s_1^2 = \frac{1}{k-1}\sum_{i=1}^{k} n_i(\overline{y}_i - \overline{y})^2 \ .$$

Som teststørrelse benyttes F-teststørrelsen

$$F = \frac{s_1^2}{s_0^2},$$

dvs. forholdet mellem variationen mellem grupper og variationen inden for grupper. – Store F-værdier er signifikante.

Testsandsynlighed: Testsandsynligheden er sandsynligheden for at få en større F-værdi i F-fordelingen med frihedsgradsantal $k-1$ og $n-k$, altså

$$\varepsilon = \mathrm{P}\left(F_{k-1,n-k} > F_{\mathrm{obs}}\right) \ ,$$

der let bestemmes ved brug af tabeller over fraktiler i F-fordelingen (i nogle tabelværker hedder F-fordelingen v^2-fordelingen).

11.4 Tosidet variansanalyse

I ensidet variansanalyse sammenligner man k grupper af normalfordelte observationer, og spørgsmålet er om grupperne er ens eller ej. Det kan siges at være en noget unuanceret problemstilling; ofte vil man være interesseret i at vide noget om på hvilken måde grupperne er forskellige når de ikke er ens. I mange situationer sidder modelbyggeren inde med noget ekstra information om grupperne, information der kan give fingerpeg om *hvordan* de er forskellige. Her skal vi nu beskæftige os med ét eksempel på, hvordan man i den statistiske model kan indbygge viden om forsøgsplanen; en anden slags eksempler præsenteres i Kapitel 12.

Tosidet variansanalyse er analysemetode man kan benytte når der er tale om normalfordelte observationer der er inddelt i grupper, og når grupperne er fastlagt ved hjælp af to forskellige kriterier, *faktorer*.

Eksempel 11.4. ***Kartoffeldyrkning***

For at opnå de optimale vækstbetingelser skal planter have de fornødne næringsstoffer i de rette forhold. Dette eksempel handler om at bestemme det rette forhold mellem mængden af tilført kvælstofgødning og mængden af tilført fosforgødning til kartofler. Man har dyrket nogle kartoffelmarker på seks forskellige måder, svarende til seks forskellige kombinationer af mængde tilført kvælstof (0, 1 eller 2 enheder) og mængde tilført fosfor (0 eller 1 enhed) og derefter målt høstudbyttet. På den måde får man nogle observationer, høstudbytterne, som er inddelt i grupper efter dyrkningsmetode, således at grupperne er fastlagt ved hjælp af to kriterier, nemlig tilført kvælstof og tilført fosfor.

Et sådant dyrkningsforsøg udført i 1932 ved Ely gav de resultater der er vist i Tabel 11.5 [10].

[10] Bemærk i øvrigt at høstudbytterne har undergået visse forandringer på deres vej til Tabel 11.5. Den væsentligste er at man har taget logaritmen til tallene. Grunden hertil er, at erfaringsmæssigt er høstudbyttet af kartofler ikke særlig normalfordelt, hvorimod det det ser bedre ud med logaritmen til høstudbyttet. Da man havde taget logaritmen til tallene viste det sig, at alle resultaterne hed '3-komma-et-eller-andet', så for at få nogle pæne tal ud af det har man trukket 3 fra og ganget med 1000.

Tabel 11.5: Udbytte ved dyrkningsforsøg med kartofler. Værdierne er 1000 × (log(udbytte målt i lbs) − 3) for 36 parceller.

		kvælstof					
		0			1		
fosfor	0	591	450	584	619	618	524
		509	636	413	651	655	564
	1	722	689	625	801	688	682
		584	614	513	703	774	623
	2	702	677	684	814	757	810
		643	668	699	792	790	703

Opgaven er nu at undersøge hvordan de to faktorer kvælstof og fosfor virker hver for sig og sammen; er det f.eks. sådan at virkningen af at gå fra en til to enheder forsfor afhænger af om der tilføres kvælstof eller ej?

Det kan undersøges med en tosidet variansanalyse-model. □

Grundmodellen

Grundmodellen er den at der foreligger nogle observationer y som er ordnet i et antal grupper, således at observationer fra samme gruppe tænkes at stamme fra samme normalfordeling, og således at alle normalfordelingerne har samme varians. Situationen er altså tilsyneladende den samme som i ensidet variansanalyse; det nye er, at grupperne nu ikke længere bare er nummereret fra 1 til k på en eller anden måde, men at de er fastlagt ved hjælp af *to* separate inddelingskriterier, *faktorer*. Man kan tænke på grupperne som anbragt i *celler* i et tosidet skema med r rækker og s søjler, hvor r og s er antallene af niveauer for de to faktorer. En gruppe bliver på denne måde indiceret ved hjælp af *to* indices, et rækkenummer i og et søjlenummer j, så man kan tale om gruppe nr. (i,j). I den generelle gennemgang af metoden betegner n_{ij} antallet af observationer i gruppe (i,j), og de enkelte observationer i denne gruppe betegnes y_{ijk}, $k = 1,2,\ldots,n_{ij}$. Det totale antal observationer betegnes n.

I Tabel 11.5 er $r = 3$ og $s = 2$, og alle n_{ij}-erne er lig 6. Generelt ser situationen sådan ud:

	$j=1$	$j=2$	...	$j=s$
$i=1$	y_{11k} , $k=1,\ldots,n_{11}$	y_{12k} , $k=1,\ldots,n_{12}$	...	y_{1sk} , $k=1,\ldots,n_{1s}$
$i=2$	y_{21k} , $k=1,\ldots,n_{21}$	y_{22k} , $k=1,\ldots,n_{22}$	...	y_{2sk} , $k=1,\ldots,n_{2s}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$i=r$	y_{r1k} , $k=1,\ldots,n_{r1}$	y_{r2k} , $k=1,\ldots,n_{r2}$	...	y_{rsk} , $k=1,\ldots,n_{rs}$

Observationsantallene n_{ij} kan principielt være hvad som helst. Det viser sig dog, at udregningerne bliver enklere og konklusionerne mere forståelige hvis n_{ij}-erne opfylder betingelsen

$$n_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n} , \tag{11.10}$$

hvor som sædvanlig $n_{i\bullet} = \sum_j n_{ij}$ og $n_{\bullet j} = \sum_i n_{ij}$. Betingelsen (11.10) er specielt opfyldt hvis alle n_{ij}-erne er lige store. I det følgende vil vi gå ud fra at (11.10) er opfyldt.

Den statistiske model for den foreliggende situation skal være den, at observationerne y_{ijk} er observerede værdier af uafhængige stokastiske variable Y_{ijk}, som alle er normalfordelte med samme varians σ^2, og hvis middelværdier kan afhænge af gruppen, kort

$$Y_{ijk} \sim \mathcal{N}(\mu_{ij}, \sigma^2) . \tag{11.11}$$

Her beskriver de rs middelværdiparametre μ_{ij} den *systematiske variation* mellem de rs grupper, og variansparameteren σ^2 beskriver den *tilfældige variation* inden for grupperne.

Denne model er som nævnt blot den gamle model (11.1) hvor parametrene (og grupperne) er navngivet på en ny måde, svarende til at observationerne er arrangeret i et tosidet skema. Derfor kan vi uden videre opskrive estimaterne over de forskellige parametre:

- gruppemiddelværdierne estimeres ved gruppegennemsnittene,

$$\hat{\mu}_{ij} = \overline{y}_{ij} ,$$

Tabel 11.6: Kartoffeldyrkning: Middeltal $\overline{y}$ og variansskøn s^2 i hver af de seks grupper, jf. Tabel 11.5.

			kvælstof 0	kvælstof 1
fosfor	0	$\overline{y}$ =	530.50	605.17
		s^2 =	7680.30	2644.57
	1		624.50	711.83
			5569.90	4248.57
	2		678.83	777.67
			474.97	1745.07

- variansen estimeres ved residualkvadratsummen divideret med antallet af frihedsgrader (= antal observationer minus antal estimerede middelværdiparametre):

$$s_0^2 = \frac{1}{n-rs}\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{n_{ij}}\left(y_{ijk}-\overline{y}_{ij}\right)^2 . \qquad (11.12)$$

Eksempel 11.5. ***Kartoffeldyrkning, fortsat***

Vi vil udregne estimater mm. i kartoffeleksemplet. Man kan være interesseret i for en ordens skyld at teste grundmodellens antagelse om varianshomogenitet, og derfor beregnes for hver af de seks grupper ikke blot gennemsnit men også variansskøn, se Tabel 11.6. Endvidere finder man den samlede kvadratsum til 111817, således at det fælles skøn over variansen inden for grupper er

$$s_0^2 = \frac{111817}{36-6} = 3727.2 .$$

Bartlett's teststørrelse (11.5) *bliver* $B_{\text{obs}} = 9.5$ *der skal sammenlignes med* χ^2*-fordelingen med* $6-1=5$ *frihedsgrader. Tabelopslag viser at der er knap 10% chance for at få en større værdi, og der således ikke noget der taler alvorligt imod antagelsen om varianshomogenitet. Vi kan derfor basere de videre undersøgelser på den formodede grundmodel.* □

Additivitetshypotesen

Grunden til at man har stillet observationerne op i et tosidet skema er, at man har en formodning om at det forholder sig på den måde, at middelværdiparameteren μ_{ij} kan fås på simpel måde ud fra en niveauparameter ξ, en parameter η_i knyttet til den pågældende række og en parameter ζ_j knyttet til den pågældende søjle.

Mere præcist er formodningen

- at man i stedet for de rs ukendte middelværdiparametre μ_{ij} kan nøjes med
  1. en parameter ξ der beskriver det generelle niveau,
  2. for hver række en *rækkeparameter* η_i der beskriver "virkningen" af den pågældende række, og
  3. for hver søjle en *søjleparameter* ζ_j der beskriver "virkningen" af den pågældende søjle,
- fordi middelværdien svarende til den (i,j)-te gruppe er $\xi+\eta_i+\zeta_j$.

Denne formodning, *additivitetshypotesen* eller *hypotesen om additivitet* mellem rækkefaktoren og søjlefaktoren, skrives i den kortfattede statistiske notation som en statistisk hypotese

$$H_1 \ : \ \mu_{ij} = \xi + \eta_i + \zeta_j \tag{11.13}$$

eller (lidt mere informativt)

$$H_1 \ : \ Y_{ijk} \sim \mathcal{N}(\xi + \eta_i + \zeta_j \, , \ \sigma^2) \ . \tag{11.14}$$

Additivitetshypotesen kaldes undertiden for hypotesen om *forsvindende vekselvirkning* (mellem række- og søjlefaktor), fordi virkningen af at ændre f.eks. rækkefaktoren fra niveau i_1 til niveau i_2 er den samme for alle værdier af søjlefaktoren (nemlig $\eta_{i_2} - \eta_{i_1}$), dvs. der er ingen vekselvirkning.

Ved første øjekast kunne det se ud som om der under additivitetshypotesen er i alt $1+r+s$ middelværdiparametre, men i realiteten er der kun $r+s-1$ parametre. Det kommer af at parametrene aldrig optræder alene men altid i et udtryk af formen $\xi+\eta_i+\zeta_j$, hvilket betyder at hvis man f.eks. lægger 5 til alle η-erne og lægger 7 til alle ζ-erne og samtidig trækker 12 fra ξ, så ændrer det ikke noget ved $\xi+\eta_i+\zeta_j$, og modellen er derfor uforandret. For at råde bod på denne ubestemthed vedtager man, at

- rækkeparametrene $\eta_1, \eta_2, \ldots, \eta_r$ (der kan siges at skulle beskrive rækkernes afvigelser fra det fælles niveau ξ) skal summere til 0 når man vægter dem med antallene af observationer i de tilsvarende rækker, altså

$$n_{1\bullet}\eta_1 + n_{2\bullet}\eta_2 + \ldots + n_{r\bullet}\eta_r = 0 ,$$

eller kortere

$$\sum_{i=1}^{r} n_{i\bullet}\eta_i = 0 , \tag{11.15}$$

- søjleparametrene $\zeta_1, \zeta_2, \ldots, \zeta_r$ (der kan siges at skulle beskrive søjlernes afvigelser fra det fælles niveau ξ) skal summere til 0 når man vægter dem med antallene af observationer i de tilsvarende søjler, altså

$$n_{\bullet 1}\zeta_1 + n_{\bullet 2}\zeta_2 + \ldots + n_{\bullet r}\zeta_s = 0 ,$$

eller kortere

$$\sum_{j=1}^{s} n_{\bullet j}\zeta_j = 0 . \tag{11.16}$$

Når man pålægger middelværdiparametrene disse to ekstra betingelser, så er der ikke nogen ubestemtheder i parametriseringen[11].

Estimation af parametrene under additivitetshypotesen

De ukendte middelværdiparametre estimeres (som altid) ved brug af maximum likelihood metoden. Inden vi opskriver likelihoodfunktionen under additivitetshypotesen H_1 er det praktisk at have likelihoodfunktionen under grundmodellen (11.11); den er

$$L(\mu_{11}, \ldots, \mu_{rs}, \sigma^2)$$
$$= \prod_{i=1}^{r}\prod_{j=1}^{s}\prod_{k=1}^{n_{ij}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\tfrac{1}{2}\frac{(y_{ijk}-\mu_{ij})^2}{\sigma^2}\right)$$

[11] Dette er ikke umiddelbart indlysende, men længere fremme finder vi entydige estimater over parametrene, og netop det at estimaterne er entydige, betyder at ubestemthederne er forsvundet.

$$= (2\pi)^{-n/2}(\sigma^2)^{-n/2}\exp\left(-\tfrac{1}{2}\frac{\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{n_{ij}}(y_{ijk}-\mu_{ij})^2}{\sigma^2}\right).$$

Når vi her erstatter μ_{ij} med $\xi+\eta_i+\zeta_j$ fås likelihoodfunktionen L_1 hørende til additivitetshypotesen H_1:

$$L_1(\xi,\eta_1,\eta_2,\ldots,\eta_r,\zeta_1,\zeta_2,\ldots,\zeta_s,\sigma^2) \tag{11.17}$$

$$= (2\pi)^{-n/2}(\sigma^2)^{-n/2}\exp\left(-\tfrac{1}{2}\frac{\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{n_{ij}}\big(y_{ijk}-(\xi+\eta_i+\zeta_j)\big)^2}{\sigma^2}\right).$$

Det fremgår heraf, at de skøn over middelværdiparametrene der maksimaliserer likelihoodfunktionen L_1 er dem der minimaliserer kvadratsummen

$$\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{n_{ij}}\big(y_{ijk}-(\xi+\eta_i+\zeta_j)\big)^2, \tag{11.18}$$

altsammen under bibetingelserne $\sum_{i=1}^{r} n_{i\cdot}\eta_i = 0$, $\sum_{j=1}^{s} n_{\cdot j}\zeta_j = 0$.

Det ville nu være muligt at løse estimationsproblemet ved hjælp af standardmetoder for "bestemmelse af ekstremum under bibetingelser for en funktion af mange variable", men man kan også tænke sig lidt om og gætte en løsning som man så viser er rigtig:

- Parameteren ξ skulle beskrive det generelle niveau, så det ville ikke være overraskende hvis ξ skulle estimeres ved det totale gennemsnit

$$\overline{y}_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{n_{ij}} y_{ijk}.$$

- Parametrene $\eta_1,\eta_2,\ldots,\eta_r$ skulle beskrive de enkelte rækkers særlige afvigelse fra det generelle niveau, så derfor er det tænkeligt at η_i skal estimeres som differensen $\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot}$ mellem gennemsnittet i den i-te række og det totale gennemsnit[12]; her er

$$\overline{y}_{i\cdot} = \frac{1}{n_{i\cdot}}\sum_{j=1}^{s}\sum_{k=1}^{n_{ij}} y_{ijk}$$

[12]Bemærk at $\widehat{\eta}_i = \overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot}$ faktisk opfylder betingelsen (11.15).

det i-te rækkegennemsnit.

- Tilsvarende er det tænkeligt at ζ_j skal estimeres som differensen $\overline{y}_{\cdot j} - \overline{y}_{\cdot\cdot}$ mellem gennemsnittet i den j-te række og det totale gennemsnit[13]; her er

$$\overline{y}_{\cdot j} = \frac{1}{n_{\cdot j}} \sum_{i=1}^{r} \sum_{k=1}^{n_{ij}} y_{ijk}$$

det j-te søjlegennemsnit.

Vi vil nu vise at disse formodninger er rigtige.

Først omskriver vi differensen $y_{ijk} - (\xi + \eta_i + \zeta_j)$ mellem den (i, j, k)-te observation og den tilsvarende teoretiske værdi, idet vi lægger adskillige velvalgte led til og trækker dem fra:

$$\begin{aligned} & y_{ijk} - (\xi + \eta_i + \zeta_j) \\ &= \left(y_{ijk} - \overline{y}_{ij}\right) \\ &\quad + \left(\overline{y}_{ij} - (\overline{y}_{i\cdot} + \overline{y}_{\cdot j} - \overline{y}_{\cdot\cdot})\right) \\ &\quad + \left((\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot}) - \eta_i\right) \\ &\quad + \left((\overline{y}_{\cdot j} - \overline{y}_{\cdot\cdot}) - \zeta_j\right) \\ &\quad + (\overline{y}_{\cdot\cdot} - \xi) \ . \end{aligned}$$

Den kvadratsum der skal minimaliseres er summen over k, j og i af kvadratet af det der står på venstre side i denne omskrivning, og derfor også summen over k, j og i af kvadratet af højresiden (!). Man omskriver kvadratet på højresiden ved hjælp af formlen for kvadratet på en femleddet størrelse[14]. Når man derpå summerer over k, j og i får man

$$\sum_{i=1}^{r} \sum_{j=1}^{s} \sum_{k=1}^{n_{ij}} \left(y_{ijk} - (\xi + \eta_i + \zeta_j)\right)^2$$

[13] Bemærk at $\widehat{\zeta}_j = \overline{y}_{\cdot j} - \overline{y}_{\cdot\cdot}$ faktisk opfylder betingelsen (11.16).

[14] Kvadratet på en femleddet størrelse giver dels kvadraterne på hvert af leddene, dels de dobbelte produkter svarende til parrene af forskellige led:

$$\begin{aligned} & (A + B + C + D + E)^2 \\ &= A^2 + B^2 + C^2 + D^2 + E^2 \\ &\quad + 2AB + 2AC + 2AD + 2AE + 2BC + 2BD + 2BE + 2CD + 2CE + 2DE \ . \end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{n_{ij}} \left(y_{ijk} - \overline{y}_{ij}\right)^2 \\
&+ \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{n_{ij}} \left(\overline{y}_{ij} - (\overline{y}_{i\cdot} + \overline{y}_{\cdot j} - \overline{y}_{\cdot\cdot})\right)^2 \\
&+ \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{n_{ij}} \left((\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot}) - \eta_i\right)^2 \\
&+ \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{n_{ij}} \left((\overline{y}_{\cdot j} - \overline{y}_{\cdot\cdot}) - \zeta_j\right)^2 \\
&+ \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{n_{ij}} (\overline{y}_{\cdot\cdot} - \xi)^2 \ ,
\end{aligned}
$$

idet det nemlig kan vises at alle de 10 dobbelte produkter bliver 0 [15]. Omskrivningen viser, at den kvadratsum der skal minimaliseres kan spaltes i en sum af fem kvadratsummer, hvoraf de to udelukkende indeholder y-er og de tre indeholder både observationer og parametre, men hvor parametrene nu er pænt fordelt ud mellem de enkelte summer. Det ses at for ét bestemt valg af parameterværdier vil de kvadratsummer hvor der optræder parametre blive 0 (og dermed minimale), nemlig når vi lader parameterværdierne have de værdier vi tidligere gættede på at de skulle have. Maksimaliseringsestimaterne er altså

$$
\begin{aligned}
\widehat{\xi} &= \overline{y}_{\cdot\cdot} \\
\widehat{\eta}_i &= \overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot} \\
\widehat{\zeta}_j &= \overline{y}_{\cdot j} - \overline{y}_{\cdot\cdot} \ .
\end{aligned}
$$

Det bemærkes at skønnet over middelværdien i den (i,j)-te celle derved bliver

$$
\begin{aligned}
\widehat{\xi} + \widehat{\eta}_i + \widehat{\zeta}_j &= \overline{y}_{\cdot\cdot} + (\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot}) + (\overline{y}_{\cdot j} - \overline{y}_{\cdot\cdot}) \\
&= \overline{y}_{i\cdot} + \overline{y}_{\cdot j} - \overline{y}_{\cdot\cdot} \ .
\end{aligned}
$$

Hvis vi i ovenstående opspaltning af kvadratsummen indsætter de *estimerede* parametre, så bliver de tre sidste summer som nævnt 0, og

[15] Selv om det er et afgørende punkt, forbigås de nærmere detaljer i beviset for denne påstand. Beviset består simpelt hen i at skrive summerne op og udnytte definitionerne på de forskellige indgående gennemsnit. Undervejs får man brug for betingelserne (11.10), (11.15) og (11.16).

vi får

$$\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{n_{ij}}\left(y_{ijk}-(\overline{y}_{i\cdot}+\overline{y}_{\cdot j}-\overline{y}_{\cdot\cdot})\right)^2$$

$$=\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{n_{ij}}\left(y_{ijk}-\overline{y}_{ij}\right)^2 \qquad (11.19)$$

$$+\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{n_{ij}}\left(\overline{y}_{ij}-(\overline{y}_{i\cdot}+\overline{y}_{\cdot j}-\overline{y}_{\cdot\cdot})\right)^2 .$$

Man kan give en beskrivelse af disse kvadratsummer:

- Kvadratsummen på venstre side beskriver *den totale variation omkring additivitetshypotesen*, dvs. de enkelte y_{ijk}-observationers variation omkring den middelværdi der er estimeret under antagelse af additivitet; da der er n observationer og der er estimeret $r+s-1$ middelværdier, har kvadratsummen $n-(r+s-1)$ frihedsgrader.

- Den første kvadratsum på højre side beskriver *variationen inden for grupper*, dvs. de enkelte y_{ijk}-observationers variation omkring den middelværdi der er estimeret i grundmodellen (dvs. omkring gruppegennemsnittet $\overline{y}_{ij}$); da der er n observationer og rs estimerede gruppemiddelværdier, har kvadratsummen $n-rs$ frihedsgrader.

- Den anden kvadratsum på højre side beskriver *vekselvirkningsvariationen*, dvs. hvordan grundmodellens estimerede middelværdier varierer omkring additivitetshypotesens estimerede middelværdier; da der er rs estimerede middelværdier i grundmodellen og $r+s-1$ estimerede middelværdier under additivitetshypotesen, har kvadratsummen $rs-(r+s-1)$ frihedsgrader.

Afslutningsvis skal vi have estimeret *variansparameteren* σ^2 under additivitetshypotesen. Maksimaliseringsestimatet for σ^2 kan bestemmes ud fra likelihoodfunktionen (11.17) hvor man indsætter de estimerede middelværdiparametre og derved står tilbage med en funktion af den ene variabel σ^2. Man finder at $\widehat{\sigma^2}$ er lig residualkvadratsummen omkring additivitetshypotesen divideret med n. I praksis benytter man

dog det skøn der fremkommer ved at dividere residualkvadratsummen med dens frihedsgradsantal, så σ^2 estimeres altså i praksis ved

$$s_{01}^2 = \frac{1}{n-(r+s-1)} \sum_{i=1}^{r} \sum_{j=1}^{s} \sum_{k=1}^{n_{ij}} \left(y_{ijk} - (\overline{y}_{i\cdot} + \overline{y}_{\cdot j} - \overline{y}_{\cdot\cdot}) \right)^2 .$$

Var den teoretiske udledning af estimaterne lang, er de udregninger der i praksis skal foretages til gengæld simple, idet man blot skal udregne forskellige gennemsnit og en sum af kvadratiske afvigelser.

Eksempel 11.6. ***Kartoffeldyrkning, fortsat***

I denne fortsættelse af eksemplet vil vi beskrive talmaterialet i Tabel 11.5 med en normalfordelingsmodel hvor virkningerne af de to faktorer "tilført fosfor" og "tilført kvælstof" indgår additivt. Vi betegner den k-te observation i den i-te række og j-te søjle y_{ijk}. Den statistiske model er den, at y_{ijk}-erne opfattes som observerede værdier af uafhængige normalfordelte stokastiske variable Y_{ijk} hvor

$$Y_{ijk} \sim \mathcal{N}(\xi + \eta_i + \zeta_j \, , \, \sigma^2) \, .$$

Her beskriver ξ det fælles niveau, η_i-erne virkningerne af tilført fosfor og ζ_j-erne virkningerne af tilført kvælstof.

Det fælles niveau ξ estimeres til

$$\overline{y}_{\cdot\cdot} = 654.75 \, ;$$

de estimerede fosforvirkninger (rækkevirkninger) er

$$\begin{aligned} \overline{y}_{1\cdot} - \overline{y}_{\cdot\cdot} &= -86.92 \\ \overline{y}_{2\cdot} - \overline{y}_{\cdot\cdot} &= 13.42 \\ \overline{y}_{3\cdot} - \overline{y}_{\cdot\cdot} &= 73.50 \, ; \end{aligned}$$

de estimerede kvælstofvirkninger (søjlevirkninger) er

$$\begin{aligned} \overline{y}_{\cdot 1} - \overline{y}_{\cdot\cdot} &= -43.47 \\ \overline{y}_{\cdot 2} - \overline{y}_{\cdot\cdot} &= 43.47 \, . \end{aligned}$$

Herudfra kan man eventuelt udregne de estimerede gruppemiddelværdier under additivitetshypotesen, se Tabel 11.7. Det va-

Tabel 11.7: Kartoffeldyrkning: Gruppegennemsnit (øverst) og estimerede gruppemiddelværdier under additivitetshypotesen (nederst).

		kvælstof 0	kvælstof 1
fosfor	0	530.50	605.17
		524.36	611.30
	1	624.50	711.83
		624.70	711.64
	2	678.83	777.67
		684.78	771.72

riansskøn der skal benyttes hvis additivitetshypotesen er rigtig er

$$s_{01}^2 = \frac{112693.5}{36-(3+2-1)} = 3521.7$$

med $36-(3+2-1)=32$ *frihedsgrader.* □

Test af additivitetshypotesen

Additivitetshypotesen kan som enhver statistisk hypotese testes ved et kvotienttest. Imidlertid kan kvotienttesstørrelsen Q omskrives på lignende måde som det skete i "Ensidet variansanalyse" (se side 214), således at man som teststørrelse får en kvotient mellem to varianser, nemlig

$$F = \frac{s_1^2}{s_0^2}, \tag{11.20}$$

hvor s_0^2 er det tidligere fundne variansskøn (11.12) i grundmodellen,

$$s_0^2 = \frac{1}{n-rs}\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{n_{ij}}\left(y_{ijk}-\overline{y}_{ij}\right)^2,$$

og s_1^2 er den såkaldte *vekselvirkningsvarians*, nemlig vekselvirkningskvadratsummen (jf. side 229) divideret med sit frihedsgradsantal:

$$s_1^2 = \frac{1}{rs-(r+s-1)} \sum_{i=1}^{r} \sum_{j=1}^{s} \sum_{k=1}^{n_{ij}} \left(\overline{y}_{ij} - (\overline{y}_{i\cdot} + \overline{y}_{\cdot j} - \overline{y}_{\cdot\cdot}) \right)^2 .$$

Store værdier af F er signifikante, og når additivitetshypotesen er rigtig, vil F følge en F-fordeling med $(rs-(r+s-1), n-rs)$ frihedsgrader. Testsandsynligheden ε findes derfor som

$$\varepsilon = \mathrm{P}\left(F_{rs-(r+s-1),\, n-rs} > F_{\mathrm{obs}} \right) .$$

Eksempel 11.7. ***Kartoffeldyrkning, fortsat***

Vi kan nu teste om der er additivitet mellem fosfor og kvælstof i kartoffeldyrkningseksemplet. Vi har tidligere fundet at

$$\begin{aligned} s_0^2 &= \frac{111817}{36-6} \\ &= 3727.2 . \end{aligned}$$

Vekselvirkningsvariansen findes til

$$\begin{aligned} s_1^2 &= \frac{877}{6-(3+2-1)} \\ &= \frac{877}{2} \\ &= 438.5 . \end{aligned}$$

Teststørrelsen er dermed

$$\begin{aligned} F &= \frac{s_1^2}{s_0^2} \\ &= \frac{438.5}{3727.2} \\ &= 0.12 , \end{aligned}$$

der skal sammenlignes med F-fordelingen med (2,30) frihedsgrader. Tabelopslag viser at testsandsynligheden er lidt under 90%,

Tabel 11.8: Kartoffel-eksemplet: *Variansanalyseskema nr. 1.*
f står for antal frihedsgrader, SS for Sum af kvadratiske afvigelser, $s^2 = SS/f$.

variation	f	SS	s^2	test
inden for grupper	30	111817	3727	
vekselvirkning	2	877	439	439/3727=0.12
additivitetshypotesen	32	112694	3522	

så der er næppe nogen tvivl om at additivitetshypotesen kan godkendes.

Som forbedret skøn over den fælles varians benyttes herefter

$$s_{01}^2 = \frac{112694}{36-(3+2-1)} = 3521.7$$

med $36-(3+2-1)=32$ *frihedsgrader.*

Man kan give en oversigt over den hidtidige del af analysen af kartoffeleksemplet i form af et variansanalyseskema, *se Tabel 11.8*
□

Undertiden kan det være en god idé at supplere med en *grafisk kontrol* af antagelsen om additivitet, ja i de tilfælde hvor der kun er én observation i hver gruppe kan man slet ikke udføre det numeriske test (da man ikke kan udregne s_0^2-størrelsen) men er under alle omstændigheder henvist til en grafisk kontrol.

Hvis additivitetshypotesen er rigtig, så gælder at de forventede værdier af gruppe- , række- og søjlegennemsnittene er[16]

$$\begin{aligned} \mathrm{E}\overline{Y}_{ij} &= \xi+\eta_i+\zeta_j\ , \\ \mathrm{E}\overline{Y}_{i\cdot} &= \xi+\eta_i\ , \\ \mathrm{E}\overline{Y}_{\cdot j} &= \xi+\zeta_j\ . \end{aligned}$$

[16]Ved udledningen af udtrykkene for $\mathrm{E}\overline{Y}_{i\cdot}$ og$\mathrm{E}\overline{Y}_{\cdot j}$ benytter man betingelserne (11.15) og (11.16).

Det betyder at hvis man for et bestemt j tænker på de r talpar $(\mathrm{E}\overline{Y}_{1\bullet}, \mathrm{E}\overline{Y}_{1j})$, $(\mathrm{E}\overline{Y}_{2\bullet}, \mathrm{E}\overline{Y}_{2j})$, $\ldots$,$(\mathrm{E}\overline{Y}_{r\bullet}, \mathrm{E}\overline{Y}_{rj})$ som punkter i et koordinatsystem, så vil disse punkter ligge på den linie som har hældning 1 og som skærer ordinataksen i ζ_j. Nu kender vi jo ikke værdierne af de teoretiske middelværdiparametre, men vi er henvist til de empiriske skøn

$$\begin{aligned} \widehat{\mathrm{E}\overline{Y}}_{ij} &= \overline{y}_{ij} \,, \\ \widehat{\mathrm{E}\overline{Y}}_{i\bullet} &= \overline{y}_{i\bullet} \,, \\ \widehat{\mathrm{E}\overline{Y}}_{\bullet j} &= \overline{y}_{\bullet j} \,. \end{aligned}$$

Hvis additivitetshypotesen er rigtig, så skal man forvente at de r punkter $(\overline{y}_{1\bullet}, \overline{y}_{1j})$, $(\overline{y}_{2\bullet}, \overline{y}_{2j})$,$\ldots$, $(\overline{y}_{r\bullet}, \overline{y}_{rj})$ ligger nogenlunde omkring en linie med hældning 1 og som skærer ordinataksen i $\overline{y}_{\bullet j} - \overline{y}_{\bullet\bullet}$. Dette skal gælde for alle s forskellige j-værdier. – Man kan naturligvis gøre noget tilsvarende hvor der er byttet om på rækker og søjler.

Signifikans af række- og søjlevirkninger

Antag at additivitetshypotesen er godtaget. Det betyder, at det oprindelige problem hvor der var rs grupper med hver sin helt egen middelværdiparameter μ_{ij} nu er simplificeret til en situation, hvor de rs gruppers middelværdier hænger sammen på en sådan måde, at der kun indgår $r+s+1$ parametre, nemlig r rækkeparametre η_i, s søjleparametre ζ_j og et fælles niveau ξ. Middelværdien i den (i,j)-te gruppe er så $\xi+\eta_i+\zeta_j$. Den aktuelle statistiske model kan derfor kortfattet resumeres som at y_{ijk}-erne er observerede værdier af uafhængige stokastiske variable Y_{ijk} hvor

$$Y_{ijk} \sim \mathcal{N}(\xi+\eta_i+\zeta_j\,,\,\sigma^2)\,. \tag{11.21}$$

η_i-erne og ζ_j-erne skal opfylde den bibetingelse at de skal summere til 0, se (11.15) og (11.16).

Det kan nu undertiden være af interesse at teste, om *søjlevirkningen* er signifikant, dvs. om der overhovedet er forskel på søjlerne. At der ikke er forskel på søjlerne kan formuleres som at de s søjleparametre ζ_j er ens, og da ζ-erne skal summere til 0 må de så alle være 0. Hypotesen om *forsvindende søjlevirkning*, dvs. hypotesen om at der ikke er nogen

signifikant forskel på søjlerne, kan derfor alt i alt udtrykkes som den statistiske hypotese

$$H_2 : \zeta_1 = \zeta_2 = \ldots = \zeta_s = 0$$

eller (lidt mere informativt)

$$H_2 : Y_{ijk} \sim \mathcal{N}(\xi + \eta_i , \sigma^2) .$$

Hvis der ikke er nogen forskel på søjlerne så behøver man altså i realiteten kun at inddele observationerne efter ét kriterium, nemlig rækkekriteriet.

Måske var man mere interesseret i at undersøge *rækkevirkningen*. Hypotesen om *forsvindende rækkevirkning* kan udtrykkes som den statistiske hypotese

$$H_2^* : \eta_1 = \eta_2 = \ldots = \eta_r = 0$$

eller (lidt mere informativt)

$$H_2^* : Y_{ijk} \sim \mathcal{N}(\xi + \eta_j , \sigma^2) .$$

Dette svarer til at man i realiteten kun behøver at inddele observationerne efter ét kriterium, søjle-kriteriet.

Hvis der hverken er forskel mellem rækker eller mellem søjler, så kan alle Y_{ijk}-erne antages at stamme fra en og samme normalfordeling.

Det er værd at understrege, at det først er under antagelse af den additive model (11.21) at det overhovedet er meningsfuldt at tale om en søjlevirkning eller en rækkevirkning og at spørge om den ene eller den anden af disse er signifikante.

Estimation af parametre under H_2 og H_2^*

Den likelihoodfunktion der er gældende under den aktuelle grundmodel (11.21) (den additive model) er

$$L_1(\xi, \eta_1, \eta_2, \ldots, \eta_r, \zeta_1, \zeta_2, \ldots, \zeta_s, \sigma^2)$$

som er defineret i (11.17) på side 226. Når man skal estimere parametrene under H_2 kan det derfor i princippet ske ved at maksimalisere funktionen

$$L_1(\xi, \eta_1, \eta_2, \ldots, \eta_r, 0, 0, \ldots, 0, \sigma^2) ,$$

men man kan spare meget arbejde ved at tænke sig om en ekstra gang. Da H_2 svarer til at observationerne er inddelt efter ét kriterium kan vi uden videre benytte resultatet på side 206, og vi får at middelværdien $\xi + \eta_i$ estimeres ved gennemsnittet i den i-te række,

$$\widehat{\xi + \eta_i} = \overline{y}_{i\cdot} , \tag{11.22}$$

og variansen σ^2 estimeres ved residualkvadratsummen divideret med antallet af frihedsgrader,

$$s_{02}^2 = \frac{1}{n-r} \sum_{i=1}^{r} \sum_{j=1}^{s} \sum_{k=1}^{n_{ij}} \left(y_{ijk} - \overline{y}_{i\cdot}\right)^2 .$$

Af (11.22) plus betingelsen om at η_i-erne skal summere til 0 får man at $\widehat{\xi} = \overline{y}_{\cdot\cdot}$ og $\widehat{\eta}_i = \overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot}$, altså nøjagtig de samme estimater som under grundmodellen (11.21).

Tilsvarende er skønnene over middelværdiparametrene under H_2^* $\widehat{\xi} = \overline{y}_{\cdot\cdot}$ og $\widehat{\zeta}_j = \overline{y}_{\cdot j} - \overline{y}_{\cdot\cdot}$, og variansskønnet er

$$s_{02}^{*\,2} = \frac{1}{n-s} \sum_{i=1}^{r} \sum_{j=1}^{s} \sum_{k=1}^{n_{ij}} \left(y_{ijk} - \overline{y}_{\cdot j}\right)^2 .$$

Test af H_2 og H_2^*

Hypoteserne H_2 og H_2^* kan i princippet testes med et kvotienttest der sammenligner to likelihoodværdier, men på samme måde som ved alle andre de andre tests for middelværdihypoteser der indtil nu har været omtalt i dette kapitel, kan kvotienttesteststørrelsen omskrives til en F-teststørrelse der sammenligner to varianser.

Først omtales teststørrelsen for H_2. Som sædvanlig skal vi have spaltet residualkvadratsummen og det tilhørende frihedsgradsantal. Af omskrivningen

$$\begin{aligned} & y_{ijk} - \overline{y}_{i\cdot} \\ &= \left(y_{ijk} - (\overline{y}_{i\cdot} + \overline{y}_{\cdot j} - \overline{y}_{\cdot\cdot})\right) + \left(\overline{y}_{\cdot j} - \overline{y}_{\cdot\cdot}\right) \end{aligned}$$

får man ved at kvadrere og summere (jf. evt. fodnote 15)

$$\sum_{i=1}^{r} \sum_{j=1}^{s} \sum_{k=1}^{n_{ij}} \left(y_{ijk} - \overline{y}_{i\cdot}\right)^2$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{n_{ij}} \left(y_{ijk} - (\overline{y}_{i\cdot} + \overline{y}_{\cdot j} - \overline{y}_{\cdot\cdot})\right)^2$$

$$+ \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{n_{ij}} \left(\overline{y}_{\cdot j} - \overline{y}_{\cdot\cdot}\right)^2 .$$

Kvadratsummerne kan beskrives således:

- Kvadratsummen på venstre side er *den totale variation omkring* H_2, dvs. residualkvadratsummen under H_2. Da der er r parametre der skal estimeres under H_2 og der i alt er n observationer, så har denne kvadratsum $n - r$ frihedsgrader.

- Den første kvadratsum på højre side er residualkvadratsummen under den aktuelle grundmodel (den additive model). Den har $n - (r + s - 1)$ frihedsgrader.

- Den anden kvadratsum på højre side beskriver *søjlevariationen*, søjlegennemsnittenes variation omkring totalgennemsnittet. Der er s søjler og 1 totalgennemsnit dermed $s - 1$ frihedsgrader. Bemærk i øvrigt at denne kvadratsum også kan skrives som

$$\sum_{j=1}^{s} n_{ij} \left(\overline{y}_{\cdot j} - \overline{y}_{\cdot\cdot}\right)^2$$

eller

$$\sum_{j=1}^{s} n_{ij} \widehat{\zeta}_j^2 .$$

Ved at dividere disse kvadratsummer med deres frihedsgradsantal får vi for det første variansskønnet s_{02}^2 under H_2, for det andet det tidligere fundne variansskøn s_{01}^2 under additivitetsmodellen der nu er den aktuelle grundmodel, og for det tredie skønnet over variansen mellem søjler

$$s_2^2 = \frac{1}{s-1}\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{n_{ij}} \left(\overline{y}_{\cdot j} - \overline{y}_{\cdot\cdot}\right)^2$$

$$= \frac{1}{s-1}\sum_{j=1}^{s} n_{\cdot j} \left(\overline{y}_{\cdot j} - \overline{y}_{\cdot\cdot}\right)^2 .$$

Testet for hypotesen H_2 om ingen forskel på søjler går da ud på at sammenligne skønnet over variansen mellem søjler med variansskønnet under den aktuelle grundmodel: Man udregner

$$F = \frac{s_2^2}{s_{01}^2} ;$$

store værdier af F er signifikante. Man kan bevise, at når H_2 er rigtig, så vil F følge en F-fordeling med $(s-1, n-(r+s-1))$ frihedsgrader. Testsandsynligheden ε for at få noget mere afvigende end det faktisk opnåede findes derfor som

$$\varepsilon = \mathrm{P}\left(F_{s-1,\,n-(r+s-1)} > F_{\mathrm{obs}}\right) .$$

Testet for hypotesen H_2^* om ingen forskel på rækker forløber på ganske tilsvarende måde. Man udregner en varians mellem rækker, nemlig

$$\begin{aligned} s_2^{*2} &= \frac{1}{r-1} \sum_{i=1}^{r} \sum_{j=1}^{s} \sum_{k=1}^{n_{ij}} \left(\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot}\right)^2 \\ &= \frac{1}{r-1} \sum_{i=1}^{r} n_{i\cdot} \left(\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot}\right)^2 . \end{aligned}$$

F-teststørrelsen er da

$$F = \frac{s_2^{*2}}{s_{01}^2}$$

og den er F-fordelt med $(r-1, n-(r+s-1))$ når hypotesen er rigtig, hvilket betyder at testsandsynligheden ε bestemmes som

$$\varepsilon = \mathrm{P}\left(F_{r-1,\,n-(r+s-1)} > F_{\mathrm{obs}}\right) .$$

Det fremhæves at begge hypoteserne H_2 og H_2^* testes i forhold til den additive model (11.21). Det skal dog tilføjes, at der ikke er nogen der siger at man *skal* teste dem begge to – her som i alle andre sammenhænge gælder, at man kun skal teste hypoteser der er fornuftige og relevante i forhold til den pågældende faglige problemstilling.

Traditionelt plejer man at opsummere resultaterne af sin variansanalyse i et *variansanalyseskema*. Tabel (11.9) er et eksempel herpå.

Tabel 11.9: Kartoffel-eksemplet: *Variansanalyseskema nr. 2.*

f står for antal frihedsgrader, SS for Sum af kvadratiske afvigelser, $s^2 = SS/f$.

variation	f	SS	s^2	test
inden for grupper	30	111817	3727	
vekselvirkning	2	877	439	439/3727=0.12
additivitetshypotesen	32	112694	3522	
mellem N-niveauer	1	68034	68034	68034/3522=19
mellem P-niveauer	2	157641	78820	78820/3522=22
omkring total-gns.	35	338369		

Eksempel 11.8. ***Kartoffeldyrkning, fortsat***

Vi fandt tidligere at virkningerne af fosfor- og kvælstofgødning kunne antages at indgå additivt i udbyttebestemmelsen og at der altså ikke er nogen signifikant vekselvirkning mellem de to gødningsarter. Det kan så være af interesse at undersøge, om det overhovedet har en signifikant virkning at tilføre dels fosfor, dels kvælstof.

For at undersøge om kvælstof har en virkning testes hypotesen H_2 om forsvindende kvælstofvirkning (søjlevirkning). Variansen mellem kvælstof-niveauer udregnes til

$$\begin{aligned} s_2^2 &= \frac{68034}{2-1} \\ &= 68034 \, . \end{aligned}$$

Variansskønnet i den additive model er tidligere fundet til $s_{02}^2 = 3522$ med 32 frihedsgrader (jf. f.eks. Tabel 11.8). F-teststørrelsen bliver dermed

$$\begin{aligned} F_{\text{obs}} &= \frac{s_2^2}{s_{02}^2} \\ &= \frac{68034}{3522} \\ &= 19 \, , \end{aligned}$$

som skal sammenlignes med $F_{1,32}$-fordelingen. Værdien $F_{\text{obs}} = 19$ er ganske utvetydigt signifikant stor, således at hypotesen om forsvindende kvælstofvirkning må forkastes, dvs. konklusionen bliver at det har *en virkning at tilføre kvælstof.*

Hvis man undersøger om der er en forsvindende fosforvirkning, så må også denne hypotese forkastes, dvs. det har også en signifikant virkning at tilføre fosforgødning. Se i øvrigt Tabel 11.9. □

Resumé 11. Tosidet variansanalyse

Situation: Der foreligger nogle observationer y som er målt på en kontinuert måleskala og som er inddelt i rs grupper, celler, ved hjælp af to kriterier, en rækkefaktor med r niveauer og en søjlefaktor med s niveauer. Skematisk ser det sådan ud:

	$j=1$	$j=2$	$\ldots$	$j=s$
$i=1$	y_{11k} , $k=1,\ldots,n_{11}$	y_{12k} , $k=1,\ldots,n_{12}$	$\ldots$	y_{1sk} , $k=1,\ldots,n_{1s}$
$i=2$	y_{21k} , $k=1,\ldots,n_{21}$	y_{22k} , $k=1,\ldots,n_{22}$	$\ldots$	y_{2sk} , $k=1,\ldots,n_{2s}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$i=r$	y_{r1k} , $k=1,\ldots,n_{r1}$	y_{r2k} , $k=1,\ldots,n_{r2}$	$\ldots$	y_{rsk} , $k=1,\ldots,n_{rs}$

I den (i,j)-te celle er der n_{ij} observationer, og den k-te af disse betegnes y_{ijk}. Det totale antal observationer er n. Observationsantallene skal opfylde betingelsen

$$n_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n} ,$$

hvor $n_{i\bullet} = \sum_j n_{ij}$ og $n_{\bullet j} = \sum_i n_{ij}$.

Model: Det antages at y_{ijk}-erne er observerede værdier af uafhængige normalfordelte stokastiske variable Y_{ijk}, således at

$$Y_{ijk} \sim \mathcal{N}(\mu_{ij}, \sigma^2) ,$$

hvor μ_{ij}-erne og σ^2 er ukendte parametre. De rs middelværdiparametre μ_{ij} beskriver den systematiske variation mellem de rs grupper, og variansparameteren σ^2 (samt normalfordelingen) beskriver den tilfældige variation inden for grupper.

Estimation: Middelværdiparameteren μ_{ij} estimeres ved gennemsnittet af observationerne i den (i,j)-te celle:

$$\begin{aligned} \hat{\mu}_{ij} &= \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} y_{ijk} \\ &= \overline{y}_{ij} . \end{aligned}$$

Variansparameteren σ^2 estimeres ved residualkvadratsummen divideret med antal frihedsgrader, dvs. ved

$$s_0^2 = \frac{1}{n-rs} \sum_{i=1}^{r} \sum_{j=1}^{s} \sum_{k=1}^{n_{ij}} \left(y_{ijk} - \overline{y}_{ij} \right)^2 .$$

Modelkontrol: Hvis der er tilstrækkelig mange observationer i grupperne kan man for hver gruppe tegne et histogram (samt den fittede normalfordelingtæthed) og/eller et fraktildiagram (samt den rette linie svarende til den fittede normalfordeling).

Antagelsen om varianshomogenitet kan eventuelt testes med Bartlett's test for varianshomogenitet.

Additivitetshypotesen: Man ønsker at teste hypotesen

$$H_1 \ : \ \mu_{ij} = \xi + \eta_i + \zeta_j$$

eller

$$H_1 \ : \ Y_{ijk} \sim \mathcal{N}(\xi + \eta_i + \zeta_j \, , \, \sigma^2)$$

om additivitet mellem række- og søjlevirkninger, hvor

- parameteren ξ beskriver det fælles niveau.
- rækkeparametrene $\eta_1, \eta_2, \ldots, \eta_r$ beskriver de enkelte rækkers afvigelser fra det fælles niveau. Der et bånd på dem, nemlig at $\sum_{i=1}^{r} n_{i\cdot}\eta_i \ = \ 0$.
- søjleparametrene $\zeta_1, \zeta_2, \ldots, \zeta_s$ beskriver de enkelte søjlers afvigelser fra det fælles niveau. Der er et bånd på dem, nemlig at $\sum_{j=1}^{s} n_{\cdot j}\zeta_j \ = \ 0$.

Estimater under H_1: Under antagelse af additivitetshypotesen gælder

- Det fælles niveau estimeres ved det totale gennemsnit:

$$\widehat{\xi} \ = \ \overline{y}_{\cdot\cdot} \, .$$

- Rækkeparametrene estimeres ved differensen mellem rækkegennemsnittene og totalgennemsnittet:

$$\widehat{\eta}_i \ = \ \overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot} \, .$$

- Søjleparametrene estimeres ved differensen mellem søjlegennemsnittene og totalgennemsnittet:

$$\widehat{\zeta}_j \ = \ \overline{y}_{\cdot j} - \overline{y}_{\cdot\cdot} \, .$$

- Estimatet over middelværdien i celle (i, j) bliver dermed

$$\widehat{\xi} + \widehat{\eta}_i + \widehat{\zeta}_j \ = \ \overline{y}_{i\cdot} + \overline{y}_{\cdot j} - \overline{y}_{\cdot\cdot} \, .$$

- Variansparameteren σ^2 estimeres ved residualkvadratsummen divideret med sit antal frihedsgrader, dvs. ved $s_{01}^2 = \frac{1}{n-(r+s-1)} \sum_{i=1}^{r} \sum_{j=1}^{s} \sum_{k=1}^{n_{ij}} \left(y_{ijk} - (\overline{y}_{i\cdot} + \overline{y}_{\cdot j} - \overline{y}_{\cdot\cdot}) \right)^2$.

Teststørrelse: Udregn vekselvirkningsvariansen $s_1^2 = \frac{1}{rs-(r+s-1)} \sum_{i=1}^{r} \sum_{j=1}^{s} \sum_{k=1}^{n_{ij}} \left(\overline{y}_{ij} - (\overline{y}_{i\cdot} + \overline{y}_{\cdot j} - \overline{y}_{\cdot\cdot}) \right)^2$. Som teststørrelse for additivitetshypotesen benyttes

$$F = \frac{s_1^2}{s_0^2} ,$$

dvs. forholdet mellem vekselvirkningsvariansen og grundmodellens variansskøn. – Store værdier af F er signifikante.

Testsandsynlighed: Testsandsynligheden ε er sandsynligheden for at få en større F-værdi end den observerede i F-fordelingen med $rs-(r+s-1)$ og $n-rs$ frihedsgrader, dvs.

$$\varepsilon = \mathrm{P}\left(F_{rs-(r+s-1),\, n-rs} > F_{\mathrm{obs}} \right) .$$

Testet kan kun udføres hvis frihedsgradsantallene er større end 0.

Grafisk kontrol af additivitetshypotesen: Man kan supplere det numeriske test med et grafisk test. I et koordinatsystem afsættes for hvert j punkterne $(\overline{y}_{1\cdot}, \overline{y}_{1j})$, $(\overline{y}_{2\cdot}, \overline{y}_{2j})$, $\ldots$, $(\overline{y}_{r\cdot}, \overline{y}_{rj})$. Hvis additivitetshypotesen er rigtig skal disse punkter fordele sig nogenlunde tilfældigt omkring en linie med hældning 1. – Man kan gøre noget tilsvarende med rækker og søjler ombyttet.

Hvis der er additivitet kan man undersøge om der er en signifikant rækkevirkning og/eller en signifikant søjlevirkning.

Test for søjlevirkninger.

- Hypotesen der testes er hypotesen

$$H_2 : \zeta_1 = \zeta_2 = \ldots = \zeta_s = 0$$

om at der ikke er signifikant forskel på de s søjler.

- Teststørrelsen er forholdet

$$F = \frac{s_2^2}{s_{01}^2}$$

mellem skønnet over variansen mellem søjler s_2^2 og den aktuelle models variansskøn s_{01}^2. Her er

$$s_2^2 = \frac{1}{s-1} \sum_{j=1}^{s} n_{\cdot j} \left(\overline{y}_{\cdot j} - \overline{y}_{\cdot\cdot} \right)^2 .$$

– Store værdier af F er signifikante.

- Testsandsynligheden ε findes let ved hjælp af F-fordelingen med frihedsgrader $s-1$ og $n-(r+s-1)$:

$$\varepsilon = \mathrm{P}\left(F_{s-1,\, n-(r+s-1)} > F_{\mathrm{obs}} \right) .$$

Test for rækkevirkninger.

- Hypotesen der testes er hypotesen

$$H_2^* : \eta_1 = \eta_2 = \ldots = \eta_r = 0$$

om at der ikke er signifikant forskel på de r rækker.

- Teststørrelsen er forholdet

$$F = \frac{s_2^{*2}}{s_{01}^2}$$

mellem skønnet over variansen mellem rækker s_2^{*2} og den aktuelle models variansskøn s_{01}^2. Her er

$$s_2^{*2} = \frac{1}{r-1} \sum_{i=1}^{r} n_{i\cdot} \left(\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot} \right)^2 .$$

– Store værdier af F er signifikante.

- Testsandsynligheden ε findes let ved hjælp af F-fordelingen med frihedsgrader $r-1$ og $n-(r+s-1)$:

$$\varepsilon \;=\; \mathrm{P}\left(F_{r-1,\,n-(r+s-1)} > F_{\mathrm{obs}}\right) .$$

11.5 Tostikprøveproblemer i normalfordelingen

De metoder der benyttes for at sammenligne k grupper kan naturligvis også benyttes når $k = 2$, men der er tradition for at man alligevel giver en særlig omtale af dette specialtilfælde. Det hænger sammen med at tostikprøve-metoderne kan udformes på en særlig enkel måde, og med at man tit har brug for at sammenligne netop *to* stikprøver.

Tidligere i dette kapitel har vi diskuteret to forskellige slags modeller til sammenligning af normalfordelte observationer, nemlig dels ensidet, dels tosidet variansanalyse, svarende til at observationerne var grupperet efter et eller to kriterier. Hver af disse modeller kan specialiseres til at handle om en tostikprøvesituation, nemlig henholdsvis tostikprøveproblemet med *uparrede* observationer og tostikprøveproblemet med *parrede* observationer.

Tostikprøveproblemet med uparrede observationer

Situationen er den at man har to grupper af "individer" og på hvert individ har man målt værdien af en bestemt variabel Y. Individerne i den ene gruppe hører ikke sammen med dem i den anden gruppe på

nogen måde, de er *uparrede*. Der behøver heller ikke at være lige mange observationer i de to grupper. Skematisk ser situationen sådan ud

gruppe	observationer
1	y_{11} y_{12} $\cdots$ y_{1j} $\cdots$ y_{1n_1}
2	y_{21} y_{22} $\cdots$ y_{2j} $\cdots$ y_{2n_2}

Her betegner y_{ij} observation nr. j i gruppe nr. i, $i = 1, 2$. Grupperne har henholdsvis n_1 og n_2 observationer. Vi vil gå ud fra, at forskellen mellem observationer inden for en gruppe er tilfældig, hvorimod der er en systematisk forskel på to de grupper. I den statistiske model går vi ud fra at y_{ij}-erne er observerede værdier af uafhængige stokastiske variable Y_{ij}, som er normalfordelte med samme varians σ^2 og med middelværdier henholdsvis μ_1 og μ_2, kort

$$\begin{aligned} Y_{1j} &\sim \mathcal{N}(\mu_1\,,\,\sigma^2) \\ Y_{2j} &\sim \mathcal{N}(\mu_2\,,\,\sigma^2)\,. \end{aligned}$$

På denne måde beskriver de to middelværdiparametre μ_1 og μ_2 den *systematiske variation*, dvs. de to gruppers niveauer, medens variansparameteren σ^2 (samt normalfordelingen) beskriver den *tilfældige variation*, der altså er den samme i begge grupper (denne antagelse kan man eventuelt teste, jf. side 210). Alt i alt er det altså blot det generelle k-stikprøveproblem fra side 204, og de tidligere resultater kan uden videre overføres til den foreliggende situation:

De to ukendte middelværdiparametre estimeres ved de tilsvarende gruppegennemsnit $\overline{y}_1$ og $\overline{y}_2$, og variansparameteren σ^2 estimeres ved

$$\begin{aligned} s_0^2 &= \frac{1}{n-2}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_i)^2 \\ &= \frac{1}{n_1+n_2-2}\left(\sum_{j=1}^{n_1}(y_{1j}-\overline{y}_1)^2+\sum_{j=1}^{n_2}(y_{2j}-\overline{y}_2)^2\right)\,, \end{aligned}$$

der har $n-2 = n_1+n_2-2$ frihedsgrader.

For at vurdere om der er en signifikant forskel på de to gruppers middelværdier testes den statistiske hypotese

$$H_0 : \mu_1 = \mu_2\,.$$

Når hypotesen H_0 er rigtig estimeres den fælles værdi af middelværdiparameteren ved det totale gennemsnit

$$\overline{y} = \frac{1}{n_1+n_2-2}\left(\sum_{j=1}^{n_1} y_{1j} + \sum_{j=1}^{n_2} y_{2j}\right) ,$$

og variansparameteren σ^2 ved

$$\begin{aligned} s_{01}^2 &= \frac{1}{n-1}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij}-\overline{y})^2 \\ &= \frac{1}{n_1+n_2-1}\left(\sum_{j=1}^{n_1}(y_{1j}-\overline{y})^2 + \sum_{j=1}^{n_2}(y_{2j}-\overline{y})^2\right) , \end{aligned}$$

med $n-1$ frihedsgrader.

Hypotesen kan testes med F-teststørrelsen[17]

$$F = \frac{s_1^2}{s_0^2}$$

(jf. (11.9)), hvor

$$\begin{aligned} s_1^2 &= \frac{1}{2-1}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(\overline{y}_i-\overline{y})^2 \\ &= n_1(\overline{y}_1-\overline{y})^2 + n_2(\overline{y}_2-\overline{y})^2 . \end{aligned}$$

Der gælder at

$$\overline{y} = \frac{n_1\overline{y}_1 + n_2\overline{y}_2}{n_1+n_2} ,$$

og indsættes dette i udtrykket for s_1^2 fås let at

$$s_1^2 = \frac{(\overline{y}_1-\overline{y}_2)^2}{\frac{1}{n_1}+\frac{1}{n_2}} ,$$

således at F-teststørrelsen også kan skrives som

$$F = \frac{(\overline{y}_1-\overline{y}_2)^2}{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)s_0^2}$$

[17]med 1 og $n-2$ frihedsgrader.

$$= \left(\frac{\overline{y}_1 - \overline{y}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) s_0^2}} \right)^2$$

$$= t^2 ,$$

hvor

$$t = \frac{\overline{y}_1 - \overline{y}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) s_0^2}}$$

er forholdet mellem differensen mellem de to middelværdiskøn (tælleren) og og skønnet over standardafvigelsen på denne differens (nævneren).

Man plejer at benytte t og ikke F som teststørrelse, fordi t er mere umiddelbart forståelig. Selve testet bliver naturligvis det samme; vi skal altid forkaste når kvotienttesstørrelsen Q er ekstremt lille, og tidligere fandt vi at Q er lille netop når F er stor, og da $F = t^2$ betyder det at Q er lille netop når $|t|$ er stor. Testsandsynligheden, dvs. sandsynligheden (under forudsætning af H_0) for at få et sæt observationer der harmonerer *dårligere* med H_0 end de foreliggende observationer gør, skal derfor bestemmes som

$$\begin{aligned} \varepsilon &= \mathrm{P}_0\left(|t| > |t_{\mathrm{obs}}|\right) \\ &= \mathrm{P}_0\left(t > |t_{\mathrm{obs}}| \text{ eller } t < -|t_{\mathrm{obs}}|\right) . \end{aligned}$$

En væsentlig del af begrundelsen for at benytte denne t-teststørrelse er, at det kan bevises, at når hypotesen er rigtig så vil t følge den såkaldte t-fordeling med $f = n - 2$ frihedsgrader[18]. Testsandsynligheden ε kan derfor uden besvær findes ved hjælp af tabeller over t-fordelingen som[19]

$$\varepsilon = 2 \times \mathrm{P}\left(t_{n-2} > |t_{\mathrm{obs}}|\right) .$$

[18]nemlig det antal frihedsgrader som variansskønnet i nævneren har.

[19]Dette test er *tosidet* fordi de ekstreme t-værdier er på begge sider af 0, og det er det man som oftest bruger. Men en sjælden gang er man i en situation hvor man er aldeles sikker på, at hvis ikke $\mu_1 = \mu_2$ så er (lad os sige) $\mu_1 < \mu_2$, den modsatte ulighed er utænkelig, og i så fald vil man kun forkaste H_0 hvis t er langt fra 0 *og negativ*. Man foretager da et *ensidet* test og udregner testsandsynligheden som $\mathrm{P}(t_{n-2} < t_{\mathrm{obs}})$.

Eksempel 11.9. *C-vitamin*

C-vitamin (ascorbinsyre) er et veldefineret kemisk stof som man sagtens kan fremstille i laboratoriet (og i industrien), og man kan jo i sin naivitet forestille sig at virkningen af det "kunstige" C-vitamin i den menneskelige organisme er præcis lige så god som virkningen af det i naturen forekommende. For at undersøge om det nu også forholder sig sådan har man foretaget et eksperiment, godt nok ikke med mennesker men med marsvin (små gnavere).

Man delte 20 nogenlunde ens marsvin op i to grupper, hvoraf den ene fik appelsinsaft, den anden en tilsvarende mængde "kunstigt" C-vitamin. Efter seks uger med denne behandling målte man længden af fortændernes odontoblaster (som er det tandbensdannende væv). Man fik da disse resultater (i hver gruppe er observationerne ordnet efter størrelse):

	8.2	9.4	9.6	9.7
appelsinsaft:	10.0	14.5	15.2	16.1
	17.6	21.5		
	4.2	5.2	5.8	6.4
kunstig C-vit.:	7.0	7.3	10.1	11.2
	11.3	11.5		

Man kan fastslå at der må være tale om et tostikprøveproblem af en slags. Karakteren af observationerne gør det ikke urimeligt at forsøge sig med en normalfordelingmodel af en slags, og det er alt i alt nærliggende at sige at der er tale om et "tostikprøveproblem med uparrede normalfordelte observationer". Vi vil analysere observationerne ved brug af denne model.

Der udregnes forskellige størrelser der er gengivet i Tabel 11.10. Den model vi vil benytte forudsætter at de to grupper har samme varians. Det kan man jo eventuelt teste (jf. side 210) ved at udregne varianskvotienten

$$\begin{aligned} R &= \frac{s^2_{\text{appelsinsaft}}}{s^2_{\text{kunstigt}}} \\ &= \frac{19.69}{7.66} \\ &= 3.0\ . \end{aligned}$$

Tabel 11.10: C-vitamin-eksemplet: nogle beregnede størrelser.

n står for antal observationer y, S for Sum af y-er, $\overline{y}$ for gennemsnit af y-er, f for antal frihedsgrader, SS for Sum af kvadratiske afvigelser ('Sum of Squared deviations'), og s^2 for variansskøn (SS/f).

gruppe	n	S	$\overline{y}$	f	SS	s^2
appelsinsaft	10	131.8	13.18	9	177.236	19.69
kunstigt C-vit.	10	80.0	8.00	9	68.960	7.66
sum	20	211.8		18	246.196	
gennemsnit			10.59			13.68

Denne værdi skal sammenholdes med F-fordelingen med $(9,9)$ frihedsgrader i et tosidet test. Tabelopslag afslører at 95%-fraktilen er 3.18 og 97.5%-fraktilen 4.03. Der er derfor mellem 5 og 10 procents chance for at få en værre R-værdi selv om hypotesen er rigtig, og på dette grundlag vil vi ikke afvise antagelsen om varianshomogenitet. Som skøn over den fælles værdi af variansparameteren har vi $s_0^2 = 13.68$ med 18 frihedsgrader.

Vi kan nu gå over til det egentlige, nemlig at teste om der er en signifikant forskel på to gruppers niveauer. t-teststørrelsen er

$$\begin{aligned} t &= \frac{13.18 - 8.0}{\sqrt{\left(\frac{1}{10} + \frac{1}{10}\right) 13.68}} \\ &= \frac{5.18}{1.65} \\ &= 3.13\,, \end{aligned}$$

og den skal sammenholdes med t-fordelingen med 18 frihedsgrader. I denne fordeling er 99.5%-fraktilen 2.878, hvoraf vi kan slutte at der er mindre end 1% chance for at få en værdi numerisk større end den foreliggende værdi $t_{\text{obs}} = 3.13$. Konklusionen bliver derfor, at der er en klart signifikant forskel på de to grupper. Som det ses af tallene består forskellen i, at den "kunstige" gruppe har mindre odontoblast-vækst end appelsin-gruppen. Kunstigt C-vitamin synes altså ikke at virke så godt som det naturlige. □

Tostikprøveproblemet med parrede observationer

Hvor vi i forrige afsnit havde at gøre med observationer der blot var inddelt i to grupper efter *ét* kriterium, så er situationen nu at observationerne hører sammen på to ledder: dels hører hver observation til en af to mulige grupper, dels hører observationerne sammen to og to: de er parrede. Typiske eksempler er målinger på nogle forsøgsdyr (eller -personer) af en bestemt variabel *før* og *efter* en behandling; de to grupper består så hhv. af målingerne før og målingerne efter, og observationerne er parrede idet man véd hvilke målinger der stammer fra hvilke individer.

Vi viser situationen skematisk:

	gruppe nr. 1	gruppe nr. 2
par nr. 1	y_{11}	y_{12}
par nr. 2	y_{21}	y_{22}
$\vdots$	$\vdots$	$\vdots$
par nr. i	y_{i1}	y_{i2}
$\vdots$	$\vdots$	$\vdots$
par nr. r	y_{r1}	y_{r2}

Med andre ord er der r observationspar, og det i-te par består af y_{i1} og y_{i2}.

Ved opbygningen af en statistisk model bør man naturligvis udnytte den information der ligger i at vi véd hvilke observationer der hører sammen. Man kan sige at der sådan set er tale om et "tosidet variansanalyse-eksempel" med to søjler og med én observation i hver celle. Vi vil derfor benytte en model hvor der er additivitet mellem parvirkning og gruppe-virkning[20] (der typisk er en behandlings-virkning). Den statistiske model kan derfor komme til at se sådan ud:

Observationerne y_{ij} antages at være observerede værdier af uafhængige normalfordelte stokastiske variable Y_{ij}, hvor (jf. (11.14) side 224)

$$Y_{ij} \sim \mathcal{N}(\xi + \eta_i + \zeta_j\,,\,\sigma^2)\,. \tag{11.23}$$

[20] Man kan ikke *teste* additivitetshypotesen med det numeriske test (11.20) side 231, fordi når der kun er én observation i hver celle kan man ikke udregne variansskønnet s_0^2.

Her beskriver parameteren ξ det generelle niveau for observationerne, par-parametrene $\eta_1, \eta_2, \ldots, \eta_r$ (der summerer til 0) beskriver hvordan de enkelte pars niveauer afviger fra det fælles niveau ξ, og gruppe-parametrene ζ_1 og ζ_2 (der ligeledes summerer til 0) beskriver virkningen af at tilhøre den ene eller den anden gruppe. Da ζ_1 og ζ_2 summerer til 0 kunne man finde på at kalde dem hhv. $-\frac{1}{2}\delta$ og $\frac{1}{2}\delta$, for på den måde kommer δ til at betyde differensen mellem gruppe 2's og gruppe 1's virkning.

Det der er interessant er, om der er en signifikant forskel på de to grupper, når vi har taget højde for en eventuel forskel mellem par, og det er præcis det samme som at spørge om δ er forskellig fra 0 i den just formulerede model. I den tosidede variansanalyses sprogbrug er det det samme som at spørge om der ikke er nogen søjlevirkning. Man kan derfor give sig til at teste hypotesen H_2 om forsvindende søjlevirkning.

Læseren kan nu som en øvelse give sig til at finde ud af, hvordan testet tager sig ud i specialtilfældet med $s = 2$. – Vi afslører straks svaret, som er uhyre simpelt:

> For hvert par skal man danne differensen $d_i = y_{i2} - y_{i1}$. Hvis y-erne kan beskrives ved den påståede model, så er d-erne uafhængige observationer fra $\mathcal{N}(\delta\,,\, 2\sigma^2)$. Man skal betragte d-erne som udgørende et "enstikprøveproblem i normalfordelingen" og i dette enstikprøveproblem teste hypotesen $\delta = 0$ med det sædvanlige t-test (Resumé 8). Dette er præcis det samme som at teste hypotesen H_2 med F-testet[21].

Grunden til at det forholder sig sådan er, at når Y_{i1} har middelværdi $\xi + \eta_i + \zeta_1$ og varians σ^2, og Y_{i2} har middelværdi $\xi + \eta_i + \zeta_2$ og varians σ^2, så har $Y_{i2} - Y_{i1}$ middelværdi $(\xi + \eta_i + \zeta_2) - (\xi + \eta_i + \zeta_1) = \zeta_2 - \zeta_1 = \delta$ og varians $\sigma^2 + \sigma^2 = 2\sigma^2$.

Eksempel 11.10. *Sovemidler*

> *Det kemiske stof* hyoscyamin hydrobromid *kan anvendes som sovemiddel. Stoffet findes imidlertid i to udgaver, d-hyoscyamin*

[21] idet $t^2 = F$.

Tabel 11.11: Antal ekstra søvntimer ved behandling med hyoscyamin hydrobromid.

person nr.	dextro-	laevo-
1	0.7	1.9
2	-1.6	0.8
3	-0.2	1.1
4	-1.2	0.1
5	-0.1	-0.1
6	3.4	4.4
7	3.7	5.5
8	0.8	1.6
9	0.0	4.6
10	2.0	3.4

hydrobromid og l-hyoscyamin hydrobromid [22]*, og man er interesseret i at finde ud af om de to udgaver er lige gode eller ej. Derfor har man udført en forsøgsrække, hvor man på 10 forsøgspersoner har bestemt stoffernes søvnforlængende virkning. I Tabel 11.11 er vist det gennemsnitlige antal ekstra søvntimer pr. nat for hver person, dels ved behandling med d-udgaven, dels ved behandling med l-udgaven af stoffet.*

Da der er tale om at man på nogle forsøgspersoner har målt effekten af først en, så en anden behandling, vil det være nærliggende at søge at analysere talmaterialet ved hjælp af en model af typen "tostikprøveproblem med parrede observationer". Derfor bestemmes differenserne mellem virkningerne af laevo- og dextroudgaven af stoffet, se Tabel 11.12.

Vi vil opfatte tallene i Tabel 11.12 som et "enstikprøveproblem i normalfordelingen", og spørgsmålet om de to stoffer virker lige godt kan da præciseres til spørgsmålet om tallenes middelværdi er signifikant forskellig fra 0. Dette kan testes som en statistisk hypotese.

Gennemsnittet af differenserne i tabellen er $\overline{d} = 1.58$ timer, *og*

[22] l = laevo = venstre, d = dextro = højre (angiver til hvilken side stoffet afbøjer polariseret lys)

Tabel 11.12: Differenser mellem l- og d-hyoscyamin hydrobromids søvnforlængende virkning.

person nr.	differens (timer)
1	1.2
2	2.4
3	1.3
4	1.3
5	0.0
6	1.0
7	1.8
8	0.8
9	4.6
10	1.4

skønnet over variansen på differenserne er $s^2 = 1.51$ timer^2 *(med 9 frihedsgrader), svarende til at den estimerede standardafvigelse er* $s = 1.23$ timer. *Den estimerede standardafvigelse på gennemsnittet er dermed* $\sqrt{s^2/n} = \sqrt{1.51/10}$ timer $= 0.39$ timer. *t-teststørrelsen er da*

$$\begin{aligned} t &= \frac{\overline{d} - 0}{\sqrt{s^2/n}} \\ &= \frac{1.58 \text{ timer}}{0.39 \text{ timer}} \\ &= 4.06 \, . \end{aligned}$$

I t-fordelingen med 9 frihedsgrader er 99.5%-fraktilen 3.25 og 99.9%-fraktilen 4.29, så testsandsynligheden ligger et sted mellem 0.2% og 1%. Der er således ganske klart signifikans, dvs. de to stoffer virker signifikant forskelligt (og som man ser er l-stoffet det mest virksomme).

•

Dette var så et eksempel på et tostikprøveproblem med parrede observationer, men hvad var der sket hvis man af vanvare var kommet til at analysere det som om der var uparrede observationer?

Den t-størrelse man så ville udregne var en anden. Tælleren var den samme, fordi differensen mellem gennemsnittene er lig gennemsnittet af differenserne. Det variansskøn der skulle benyttes i nævneren er skønnet over den fælles varians i de to grupper, hvilket udregnes til $s_0^2 = 3.605 \text{ timer}^2$ *med 18 frihedsgrader, og teststørrelsen ville derfor blive*

$$\begin{aligned} t &= \frac{1.58 \text{ timer}}{\sqrt{\left(\frac{1}{10} + \frac{1}{10}\right) 3.605 \text{ timer}}} \\ &= \frac{1.58}{0.85} \\ &= 1.86 \, . \end{aligned}$$

Denne gang ville vi få 18 frihedsgrader i t-fordelingen, og det vil sige at 95%-fraktilen er 1.73 og 97.5%-fraktilen 2.10. Der ville altså være et sted mellem 5% og 10% chance for at få en mere ekstrem t-størrelse end 1.86, og man vil derfor almindeligvis sige at $t_{\text{obs}} = 1.86$ *ikke er signifikant stor. Dette test ville således ikke vise nogen signifikant forskel på de to stoffer.*

Grunden til at de to analyser giver forskellige resultater er, at der er en temmelig stor forskel på forsøgspersonerne:

- *I den først benyttede model (parrede observationer) er der personparametrene* $\eta_1, \eta_2, \ldots, \eta_{10}$ *til at beskrive denne forskel* [23]*, for så vidt der ikke er nogen vekselvirkning mellem sovemidler og personer. Prisen man betaler for at have en parameter for hver person er at variansskønnet kun får 9 frihedsgrader.*
- *I den anden model (uparrede observationer) skal al variationen mellem personer beskrives af variansparameteren (fordi forskellen mellem personer i denne omgang udelukkende anses for tilfældig), og til gengæld får variansskønnet hele 18 frihedsgrader. – På den anden side betyder det, at hvis der er stor forskel mellem personer, så bliver variansskønnet også stort.*

• •

[23] Selv om personparametrene tilsyneladende forsvinder ud af analysen, så har de gjort deres gavn.

Datamaterialet til dette eksempel er meget berømt (blandt statistikere), fordi det blev benyttet til et illustrativt eksempel i dén artikel[24] *hvor t-testet (i enstikprøveproblemet) blev introduceret. Artiklen er skrevet af t-testets opfinder og en af den statistiske videnskabs grundlæggere W.S. Gosset, som arbejdede som biometriker ved Guinness-bryggerierne, og som benyttede 'Student' som sit* nom de plume. □

[24] 'Student' (1908). The Probable Error of a Mean. *Biometrika* **6**, 1 - 25.

Resumé 12. Tostikprøveproblemet i normalfordelingen, uparrede observationer

Situation: Der foreligger nogle observationer y som er målt på en kontinuert måleskala og som er inddelt i to grupper:

gruppe	observationer
1	y_{11} y_{12} $\cdots$ y_{1j} $\cdots$ y_{1n_1}
2	y_{21} y_{22} $\cdots$ y_{2j} $\cdots$ y_{2n_2}

Her betegner y_{ij} observation nr. j i gruppe nr. i, $i = 1, 2$. Grupperne har henholdsvis n_1 og n_2 observationer.

Model: Det antages at y_{ij}-erne er observerede værdier af uafhængige stokastiske variable Y_{ij}, som er normalfordelte med samme varians σ^2 og med middelværdier henholdsvis μ_1 og μ_2, kort

$$\begin{aligned} Y_{1j} &\sim \mathcal{N}(\mu_1\,,\,\sigma^2) \\ Y_{2j} &\sim \mathcal{N}(\mu_2\,,\,\sigma^2)\,. \end{aligned}$$

Her beskriver de to middelværdiparametre μ_1 og μ_2 den systematiske variation, dvs. de to gruppers niveauer, medens variansparameteren σ^2 (samt af normalfordelingen) beskriver den tilfældige variation.

Estimation: Middelværdiparametrene μ_1 og μ_2 estimeres ved de tilsvarende gruppegennemsnit $\overline{y}_1$ og $\overline{y}_2$.

Variansparameteren σ^2 estimeres ved

$$s_0^2 = \frac{1}{n_1+n_2-2}\left(\sum_{j=1}^{n_1}(y_{1j}-\overline{y}_1)^2+\sum_{j=1}^{n_2}(y_{2j}-\overline{y}_2)^2\right)\,,$$

med $n-2 = n_1+n_2-2$ frihedsgrader.

Modelkontrol: Hvis der er tilstrækkelig mange observationer i grupperne kan man for hver gruppe tegne et histogram (samt den fittede normalfordelingtæthed) og/eller et fraktildiagram (samt den rette linie svarende til den fittede normalfordeling).

Antagelsen om varianshomogenitet kan eventuelt testes med Bartlett's test for varianshomogenitet (tilfældet $k = 2$).

Hypotese: Man ønsker at teste den statistiske hypotese

$$H_0 : \mu_1 = \mu_2$$

om at de to grupper har samme middelværdiparameter.

Teststørrelse: Som teststørrelse benyttes

$$t = \frac{\overline{y}_1 - \overline{y}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) s_0^2}} .$$

Store værdier af $|t|$ er signifikante.

Testsandsynlighed: Testsandsynligheden ε bestemmes som sandsynligheden for at få en værdi uden for intervallet med endepunkter $-t_{\text{obs}}$ og t_{obs} i t-fordelingen med $n-2$ frihedsgrader,

$$\varepsilon = 2 \times \mathrm{P}\left(t_{n-2} > |t_{\text{obs}}|\right) .$$

Konklusion: Hvis ε er meget lille, så er der en signifikant forskel mellem de to grupper, dvs. hypotesen forkastes. Hvis ε ikke er meget lille, kan man ikke på det foreliggende grundlag forkaste hyptesen.

Resumé 13. Tostikprøveproblemet i normalfordelingen, parrede observationer

Situation: Der foreligger r par af uafhængige observationer y som er målt på en kontinuert måleskala. Parrene er inddelt i to grupper:

	gruppe nr. 1	gruppe nr. 2
par nr. 1	y_{11}	y_{12}
par nr. 2	y_{21}	y_{22}
$\vdots$	$\vdots$	$\vdots$
par nr. i	y_{i1}	y_{i2}
$\vdots$	$\vdots$	$\vdots$
par nr. r	y_{r1}	y_{r2}

Model:

A. Det antages at y_{ij}-erne er observerede værdier af uafhængige normalfordelte stokastiske variable Y_{ij}, hvor

$$\begin{aligned} Y_{i1} &\sim \mathcal{N}(\xi + \eta_i - \tfrac{1}{2}\delta\,,\,\sigma_A^2) \\ Y_{i2} &\sim \mathcal{N}(\xi + \eta_i + \tfrac{1}{2}\delta\,,\,\sigma_A^2)\,. \end{aligned}$$

Her beskriver middelværdiparametrene ξ, $\eta_1, \eta_2, \ldots, \eta_r$ og δ den systematiske variation: ξ er det generelle niveau, $\eta_1, \eta_2, \ldots, \eta_r$ (der summerer til 0) er de enkelte pars afvigelser fra det generelle niveau, og δ er differensen mellem andet og første element i et par. Den tilfældige variation beskrives af variansparameteren σ_A^2 (samt af normalfordelingen).

B. Faktisk behøver man kun at antage følgende mere generelle model: Det antages at y_{ij}-erne er observerede værdier af stokastiske variable Y_{ij} med den egenskab at differenserne $Y_{i2} - Y_{i1}$ er uafhængige identisk normalfordelte:

$$Y_{i2} - Y_{i1} \sim \mathcal{N}(\delta\,,\,\sigma_B^2)\,,$$

hvor δ og σ_B^2 er ukendte parametre.
Der gælder at Model A medfører Model B, og i givet fald er $\sigma_B^2 = 2\sigma_A^2$.

Estimation: Kald differensen mellem elementerne i par nr. i for d_i:

$$d_i = y_{i2} - y_{i1}$$

Parameteren δ estimeres ved gennemsnittet af differenserne eller differensen af gennemsnittet:

$$\begin{aligned} \widehat{\delta} &= \overline{y}_2 - \overline{y}_1 \\ &= \overline{d} \, . \end{aligned}$$

Variansparameteren σ_B^2 i Model B estimeres ved

$$s_B^2 = \frac{1}{r-1} \sum_{i=1}^{r} \left(d_i - \overline{d} \right)^2$$

med $r-1$ frihedsgrader. (Variansparameteren σ_A^2 estimeres ved $\frac{1}{2} s_B^2$.)

Modelkontrol: For at checke om differenserne er normalfordelte kan man (hvis r er tilstrækkelig stor) over differenserne $d_1, d_2, \ldots, d_r$ tegne et histogram (samt den fittede normalfordelingtæthed) og/eller et fraktildiagram (samt den rette linie svarende til den fittede normalfordeling).

Hypotese: Man ønsker at teste den statistiske hypotese

$$H_0 \, : \, \delta = 0$$

om at der ikke er nogen signifikant forskel mellem de to elementer i et par.

Teststørrelse: Som teststørrelse benyttes

$$t = \frac{\overline{d}}{\sqrt{s_B^2 / r}} \, .$$

Store værdier af $|t|$ er signifikante.

Testsandsynlighed: Testsandsynligheden ε bestemmes som sandsynligheden for at få en værdi uden for intervallet med endepunkter $-t_{\text{obs}}$ og t_{obs} i t-fordelingen med $r-1$ frihedsgrader,

$$\varepsilon = 2 \times \mathrm{P} \left(t_{r-1} > |t_{\text{obs}}| \right) \, .$$

Konklusion: Hvis ε er meget lille, så er der en signifikant forskel mellem de to grupper, dvs. hypotesen forkastes. Hvis ε ikke er meget lille, kan man ikke på det foreliggende grundlag forkaste hyptesen.

Kapitel 12

Regressionsanalyse af normalfordelte observationer

Kunsten ved at opbygge en statistisk model for et datamateriale består blandt andet i at få indbygget den rette mængde information om forsøgsomstændighederne i modellen. Hvis modellen indeholder for lidt information om forsøgsomstændighederne kan resultatet blive, at for megen variation skal beskrives som tilfældig, således at modellens beskrivelsesevne ikke er så god som den kunne blive[1]. Omvendt bør man heller ikke indbygge *for* megen struktur i modellen – i sidste ende kunne man jo så opnå et perfekt fit – for en af pointerne ved statistiske modeller er, at det der alligevel ikke er signifikant skal man ikke modellere på anden måde end ved at putte det ind under "tilfældig variation".

Kapitel 11 handler om nogle typer modeller, variansanalysemodeller, hvor forsøgsomstændighederne består i at en eller to faktorer[2] har varieret mellem et mindre antal niveauer – derved kan man tage hensyn til *kvalitative* størrelser ved opbygningen af modeller for kontinuerte (og kvantitative) variable.

I det følgende skal vi beskæftige os med en type modeller der kan anvendes, når man på hvert af nogle "individer" har målt værdien af en bestemt kontinuert variabel y og desforuden har registreret værdien af forskellige *kvantitative* størrelser $x_1, x_2, \ldots, x_p$.

[1] Eksempel 11.10 giver et eksempel herpå.

[2] En *faktor* er en kvalitativ størrelse der benyttes til at definere grupper udfra.

Man har forskellige betegnelser for hhv. y og $x_1, x_2, \ldots, x_p$, betegnelser der antyder noget af hvad meningen er:

y kaldes	$x_1, x_2, \ldots, x_p$ kaldes
responsvariabel	
den modellerede variabel	baggrundsvariable
den afhængige variabel	de uafhængige variable
den forklarede variabel	de forklarende variable

Vi kan skitsere forskellige typer af eksempler for at antyde hvordan baggrundsvariable kan komme ind i billedet.

1. Lægen observerer den tid y som patienten overlever efter at være blevet behandlet for sygdommen, men lægen har også registreret en mængde baggrundsoplysninger om patienten, så som køn, alder, vægt, detaljer om sygdommen osv. Nogle af baggrundsoplysningerne kan måske indeholde information om hvor længe patienten kan forventes at overleve.

2. I en række nogenlunde ens i-lande har man bestemt mål for lungekræftforekomst, cigaretforbrug og forbrug af fossilt brændstof, altsammen pr. indbygger. Man kan da udnævne "lungekræftforekomst" til y-variabel og søge at "forklare" den ved hjælp af de to andre variable, der så får rollen som forklarende variable.

3. Man ønsker at undersøge et bestemt stofs giftighed. Derfor giver man det i forskellige koncentrationer til nogle grupper af forsøgsdyr og ser hvor mange af dyrene der dør. Her er koncentrationen x en uafhængig variabel som eksperimentator bestemmer værdien af, og antallet y af døde er den afhængige variabel.

En *statistisk model* i den slags situationer skal blandt andet

- udtrykke middelværdien af y-variablen som en simpel og "pæn" funktion af x-variablene, og

- angive hvilken sandsynlighedsfordeling der skal beskrive y-ernes tilfældige variation.

Dette kapitel beskæftiger sig fortrinsvis med modeller hvor den tilfældige variation beskrives af en *normalfordeling* og hvor middelværdien er en *lineær* funktion af x-variablene. Den slags modeller kan generelt

formuleres på følgende måde: For hvert individ i $(= 1, 2 \ldots, n)$ foreligger der dels en måling af en størrelse y (på en kontinuert målestkala), dels værdier af p baggrundsvariable $x_1, x_2, \ldots, x_p$. For hvert i har man altså de $p + 1$ tal

$$y_i,\ x_{i1}, x_{i2}, \ldots, x_{ik}\ ,$$

hvor y_i betegner den værdi af y der er målt på det i-te individ og x_{ij} betegner værdien af den j-te baggrundsvaribel hos individ nr. i. Modellen er da, at tallene $y_1, y_2, \ldots, y_n$ opfattes som observerede værdier af uafhængige normalfordelte stokastiske variable $Y_1, Y_2, \ldots, Y_n$, hvor

$$\begin{aligned} Y_i &\sim \mathcal{N}(\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j\ ,\ \sigma^2) \\ &= \mathcal{N}(\beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ik}\beta_p\ ,\ \sigma^2)\ . \end{aligned}$$

Her er koefficienterne $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ ukendte parametre der fastlægger hvordan middelværdien bestemmes kvantitativt ud fra baggrundsvariablene, og variansparameteren σ^2 beskriver den tilfældige variation omkring middelværdien.

Den formulerede model omtales nærmere i afsnittet om *multipel lineær regressionsanalyse*, men vi skal først og fremmest undersøge det vigtige specialtilfælde *simpel lineær regressionsanalyse*, hvor der kun er én baggrundsvariabel.

Under alle omstændigheder er der tale om *lineær* regressionsanalyse, hvilket betyder at baggrundsvariablene indgår *lineært*[3] i udtrykket for middelværdien. Det giver selvfølgelig en vis begrænsning i, hvor generelle man kan lave denne type modeller, men på den anden side kan man vælge baggrundsvariablene helt frit, og det er også tilladt at danne nye baggrundsvariable ud fra gamle[4].

[3] eller rettere: affint

[4] Hvis man f.eks. har én "naturligt givet" baggrundsvariabel t som er en nærmere fastlagt tidsstørrelse, kan man evt. indføre en ny baggrundsvariabel t^2, således at man alt i alt får den lineære regressionsmodel $\mathrm{E}\,Y_i = \beta_0 + \beta_1 t + \beta_2 t^2$.

12.1 Simpel lineær regressionsanalyse

Dette afsnit handler om situationer hvor der foreligger et antal talpar (x, y) – typisk stammer hvert talpar fra et bestemt "individ" – og hvor man ønsker at opstille en statistisk model for y-erne. Med andre ord er det y-erne der skal antages at være observerede værdier af stokastiske variable; den rolle x har er, at middelværdien af Y kan være en funktion af x. Skematisk ser det sådan ud:

observation	baggrundsvariabel
y_1	x_1
y_2	x_2
$\vdots$	$\vdots$
y_n	x_n

(12.1)

Der er n talpar, og parret hørende til "individ" nr. i betegnes (x_i, y_i). Vi formulerer en statistisk for y-erne på følgende måde:

- $y_1, y_2, \ldots, y_n$ er observerede værdier af stokastiske variable $Y_1, Y_2, \ldots, Y_n$.
- De stokastiske variable $Y_1, Y_2, \ldots, Y_n$ er uafhængige og normalfordelte med samme varians σ^2.
- $x_1, x_2, \ldots, x_n$ betragtes som faste tal – de er altså *ikke* (i denne model) observerede værdier af stokastiske variable.
- Middelværdien af Y-erne afhænger lineært[5] af x-erne, således at forstå at der findes to parametre α og β så

$$\mathrm{E}\,Y_i \;=\; \alpha + \beta x_i$$

for alle i.

Denne model plejer man at skrive kortfattet som

$$\mathrm{E}\,Y_i \;=\; \alpha + \beta x_i$$

eller lidt mere informativt som

$$Y_i \;\sim\; \mathcal{N}(\alpha + \beta x_i\,,\, \sigma^2)\,. \tag{12.2}$$

[5] eller rettere: affint

Modellen kaldes en *simpel lineær regressionsanalyse*-model. Det ses at den beskriver y-ernes *systematiske variation* ved hjælp af de to ukendte middelværdiparametre α og β samt de kendte konstanter $x_1, x_2, \ldots, x_n$. Den *tilfældige variation* beskrives ved hjælp af normalfordelingen og den ukendte variansparameter σ^2.

De to størrelser x og y indgår på vidt forskellig måde i modellen, og det er derfor ikke ligegyldigt hvad man lader være x og hvad y. I nogle tilfælde er det ganske klart hvad der er "observation" og hvad der er "baggrundsvariabel", men i andre tilfælde er det i høj grad et valg man træffer. Eksemplerne 12.1 og 12.2 illustrerer de to muligheder.

Eksempel 12.1. ***Kvælning af hunde***

Man ved, at hypoxi *(nedsat ilttilførsel til hjernen) kan bevirke at der dannes forskellige skadelige stoffer i hjernen, og at det kan medføre hjerneskader. (Hypoxien kan bl.a. forekomme ved fødsler.) Man er derfor interesseret i at udvikle en simpel metode til at afgøre* om *der har være hypoxi, og i givet fald hvor længe. Man har udført nogle forsøg for at undersøge, om koncentrationen af* hypoxantin *i cerebrospinalvæsken kan benyttes som hypoxi-indikator.*

Syv hunde er (under bedøvelse) blevet udsat for iltmangel ved sammenpresning af luftrøret, og hypoxantin-koncentrationen måltes efter 0, 6, 12 og 18 minutters forløb. Det var af forskellige grunde ikke muligt at foretage målinger på alle syv hunde til alle fire tidspunkter, og det kan heller ikke afgøres hvordan målinger og hunde hører sammen. Resultaterne af forsøget er vist i Tabel 12.1.

Man kan anskue situationen på den måde at der foreligger $n = 25$ par sammenhørende værdier af koncentration og varighed. Varighederne er kendte størrelser – de indgår i forsøgsplanen – hvorimod koncentrationerne kan betragtes som observerede værdier af stokastiske variable: tallene er ikke ens fordi der er en vis biologisk variation og en vis forsøgsusikkerhed, og det kan passende modelleres som tilfældig variation. Det er derfor nærliggende at søge at modellere tallene ved hjælp af en regressionsmodel med koncentration som y-variabel og varighed som x-variabel.

Tabel 12.1: Hypoxantin-målinger til de forskellige tidspunkter. I hver gruppe er observationerne ordnet efter størrelse.

varighed (min)	koncentration (μmol/l)						
0	0.0	0.0	1.2	1.8	2.1	2.1	3.0
6	3.0	4.9	5.1	5.1	7.0	7.9	
12	4.9	6.0	6.5	8.0	12.0		
18	9.5	10.1	12.0	12.0	13.0	16.0	17.1

Man kan naturligvis ikke på forhånd vide om varigheden i sig selv er en hensigtsmæssig x-variabel. Måske viser det sig at man bedre kan beskrive koncentrationen som en lineær funktion af logaritmen *til varigheden end som en lineær funktion af selve varigheden, men det betyder blot at der er tale om en lineær regressionsmodel med logaritmen til varigheden som x-variabel.* □

Eksempel 12.2. ***Fædre og sønner***

I slutningen af 1800-tallet opstod i England faget biometri, *et fag i grænseområdet mellem (hvad vi i vore dage forstår ved) statistik og biologi. De emner biometrikerne tog op var i høj grad emner med forbindelse til den nye og kontroversielle arvelighedslære, idet de håbede at kunne finde bekræftelser på og numeriske beskrivelser af evolutionsteorien*[6]*. Desuden var nogle af biometrikerne meget optaget af den almindelige debat om de sociale problemer i samfundet (og de var store), og de måtte derfor gøre sig overvejelser over, hvad arvelighedslæren kunne fortælle om samfundets udvikling.*

Biometrikeren F. Galton (1822 - 1911) spekulerede over det tilsyneladende almindelige forfald: hvordan kunne det være at fremragende fædre ikke fik tilsvarende fremragende sønner (– eller var det bare noget man syntes?). Nu er det vanskeligt at finde et mål for "fremragende-hed", så Galton gav sig til at undersøge

[6] de ville kort sagt matematificere evolutionsteorien ...

højde i stedet. Han foranstaltede en større indsamling af data[7] *om medlemmer af britiske familier.*

Galton foretog det vi nutildags kalder en regressionsanalyse og fandt, at høje fædre gennemsnitligt fik sønner der ikke var så høje som de selv men dog lå over gennemsnittet i befolkningen. Omvendt fik små fædre gennemsnitligt sønner der var højere end de selv men dog lå under gennemsnittet i befolkningen. Denne tilsyneladende nærmen sig det gennemsnitlige så Galton som en tilbagegang og kaldte det derfor en **regression**[8] [9].

I Tabel 12.2 er gengivet et talmateriale som to andre biometrikere indsamlede, idet de for 1078 par af far og søn registrerede faderens højde og sønnens højde. Tabellen skal læses på den måde, at der f.eks. var syv tilfælde ud af de 1078 hvor faderen var 67 inches *og sønnen 65* inches.

Der er tale om en situation med $n = 1078$ talpar (x, y), men det er ikke uden videre klart at den ene af de to højder er en "baggrundsvariabel" og den anden en "observation", faktisk må man vel sige at de er "observationer" begge to. Alligevel kan man vælge at opfatte f.eks. faderens højde som "baggrundsvariabel" og sønnens højde som "observation" og så foretage en såkaldt "regression af sønnens højde på faderens højde"; det kan man vælge at gøre hvis man er interesseret i at undersøge hvordan man kan forudsige, prædiktere, sønnens højde ud fra faderens. □

Der melder sig nu forskellige typer af spørgsmål:

1. Hvad er de bedste skøn over de indgående parametre α, β og σ^2 ?

2. Hvordan vurderer man om modellen (12.2) giver en rimelig beskrivelse af datamaterialet ?

3. Hvordan tester man hypoteser om parametrene ?

---

[7] foruden højde bl.a. også øjenfarve, temperament, kunstneriske evner, sygdomme, valg af ægtefælle og frugtbarhed.

[8] *regression* betyder *tilbagegang*.

[9] Vi kan altså takke Galton for betegnelsen regressionsanalyse. Det er vistnok også ham der skal have æren for at have udbredt betegnelsen *normalfordelingen* om normalfordelingen.

Tabel 12.2: Fordelingen af 1078 par af far og søn efter faderens højde og sønnens højde. Højderne er angivet i inches.

		Faderens højde																
		59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75
	60	-	-	-	-	1	-	1	-	-	-	-	-	-	-	-	-	-
	61	-	-	-	-	1	-	-	-	1	-	-	-	-	-	-	-	-
	62	-	1	-	-	-	1	-	-	1	-	-	-	-	-	-	-	-
S	63	-	-	-	2	2	2	4	5	3	1	-	-	1	-	-	-	-
ø	64	1	-	2	4	3	4	8	9	3	1	2	1	1	-	-	-	-
n	65	2	1	-	2	3	10	13	11	7	6	4	2	-	-	-	-	-
n	66	-	-	1	2	5	9	10	17	18	16	5	2	3	1	-	-	-
e	67	-	2	2	5	3	14	20	26	26	19	13	14	3	-	1	-	-
n	68	-	-	2	2	8	10	10	24	31	24	30	13	8	10	2	-	-
s	69	-	-	1	-	5	5	13	18	16	24	29	22	10	4	2	-	1
	70	-	-	-	-	1	3	6	19	12	20	22	19	14	6	3	2	1
h	71	-	-	-	-	-	3	5	9	10	19	15	21	11	8	5	1	1
ø	72	-	-	-	-	-	-	3	1	7	8	11	11	10	9	3	-	-
j	73	-	-	-	-	-	-	1	1	2	8	6	6	8	6	3	-	1
d	74	-	-	-	-	1	-	2	2	-	5	2	3	6	3	3	-	2
e	75	-	-	-	-	-	-	-	-	-	1	2	-	2	1	2	1	-
	76	-	-	-	-	-	-	-	-	-	1	-	-	1	1	1	-	-
	77	-	-	-	-	-	-	-	-	-	1	-	1	-	-	2	-	-
	78	-	-	-	-	-	-	-	-	-	-	1	1	-	-	1	-	-

Besvarelsen af disse spørgsmål inddrager ikke nye statistiske principper og ideer. Som altid i normalfordelingsmodeller estimerer vi middelværdiparametre ved maximum likelihood metoden mundende ud i en minimalisering af en sum af kvadratiske afvigelser[10], og som altid i normalfordelingsmodeller estimerer vi variansparameteren som residualkvadratsummen divideret med antallet af frihedsgrader. Statistiske hypoteser testes ved et likelihoodkvotienttest, og på samme måde som i Kapitel 11 kan Q-størrelsen omformes til en F-teststørrelse (og nogle gange til en t-teststørrelse).

Estimation af parametrene

De ukendte middelværdiparametre α og β skal estimeres ved at maksimalisere den til grundmodellen (12.2) hørende likelihoodfunktion

$$\begin{aligned} & L(\alpha, \beta, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\tfrac{1}{2} \frac{\big(y_i - (\alpha + \beta x_i)\big)^2}{\sigma^2} \right) \\ &= (2\pi)^{-n/2} (\sigma^2)^{-n/2} \exp\left(-\tfrac{1}{2} \frac{\sum_{i=1}^{n} \big(y_i - (\alpha + \beta x_i)\big)^2}{\sigma^2} \right) . \end{aligned}$$

Det fremgår heraf, at de bedste skøn over α og β er de værdier der minimaliserer kvadratsummen

$$\sum_{i=1}^{n} \big(y_i - (\alpha + \beta x_i)\big)^2 . \tag{12.3}$$

Disse værdier kan man enten bestemme ved hjælp af standardmetoder til bestemmelse af ekstremumspunkter for funktioner af to variable, eller man kan (som der er flere eksempler på i Kapitel 11) foretage snedige opspaltninger af kvadratsummen for at gøre det hele lettere. Her forsøger vi det sidste.

Det er hensigtsmæssigt at operere med x-ernes og y-ernes afvigelser fra deres gennemsnit $\overline{x}$ og $\overline{y}$. Derfor omskrives kvadratsummen (12.3)

[10] så man kan også kalde estimationsmetoden for *mindste kvadraters metode*.

til

$$\begin{aligned}\sum_{i=1}^{n}\Big((y_i-\overline{y})+(\overline{y}-(\alpha+\beta\overline{x}))-\beta(x_i-\overline{x})\Big)^2 \\ = \sum_{i=1}^{n}(y_i-\overline{y})^2 \qquad (12.4)\\ +n\Big(\overline{y}-(\alpha+\beta\overline{x})\Big)^2 \\ +\beta^2\sum_{i=1}^{n}(x_i-\overline{x})^2 \\ -2\beta\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})\ ,\end{aligned}$$

idet de øvrige to dobbelte produkter fra kvadreringen af den treleddede størrelse bliver 0. Omskrivningen har ført til et udtryk hvor α kun optræder i ét led, nemlig $n(\overline{y}-(\alpha+\beta\overline{x}))^2$, og det antager sin mindsteværdi 0 netop når α er lig $\overline{y}-\beta\overline{x}$. Vi mangler så kun at bestemme den bedste β-værdi, hvilket er den der minimaliserer de tre øvrige led, der er

$$\beta^2\sum_{i=1}^{n}(x_i-\overline{x})^2-2\beta\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})+\sum_{i=1}^{n}(y_i-\overline{y})^2$$

eller kort

$$\beta^2 SS_x-2\beta SP_{xy}+SS_y\ ,$$

hvor vi har benyttet de ofte anvendte betegnelser SS_x hhv. SS_y for sum af kvadratiske afvigelser[11] af x-er hhv. y-er, og SP_{xy} for sum af produkter af afvigelser af x-er og y-er.

Udtrykket $\beta^2 SS_x-2\beta SP_{xy}+SS_y$ er en andengradsfunktion af β, og da koefficienten til β^2 er positiv har funktionen ét minimumspunkt, og det findes ved at differentiere og sætte den afledede lig 0. Man får da at det bedste valg af β er

$$\begin{aligned}\widehat{\beta} &= \frac{SP_{xy}}{SS_x}\\ &= \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\ .\end{aligned}$$

[11] SS = Sum of Squared deviations

Ifølge betragtningerne ovenfor er det dertil svarende bedste valg af α

$$\widehat{\alpha} = \overline{y} - \widehat{\beta}\,\overline{x} .$$

Hermed har vi løst estimationsproblemet for så vidt angår middelværdiparametrene.

Undertiden, især når man skal udføre beregningerne mere eller mindre med håndkraft, kan man have fornøjelse af et andet udtryk for $\widehat{\beta}$ eller måske snarere for SP_{xy} og SS_x. Ved almindelige og lette formelmanipulationer finder man følgende formler, hvor hver gang det første lighedstegn er definitionslighedstegnet og det andet viser det alternative udtryk:

$$\begin{aligned} SP_{xy} &= \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \\ &= \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) \end{aligned} \tag{12.5}$$

og

$$\begin{aligned} SS_x &= \sum_{i=1}^{n}(x_i - \overline{x})^2 \\ &= \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 . \end{aligned} \tag{12.6}$$

I analogi med sidstnævnte formel gælder naturligvis også

$$\begin{aligned} SS_y &= \sum_{i=1}^{n}(y_i - \overline{y})^2 \\ &= \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2 . \end{aligned} \tag{12.7}$$

Variansskønnet er som altid residualkvadratsummen divideret med antallet af frihedsgrader. Residualkvadratsummen får vi ved at erstatte α og β med (udtrykkene for) $\widehat{\alpha}$ og $\widehat{\beta}$ i (12.3), så den er

$$\sum_{i=1}^{n}\left(y_i - (\widehat{\alpha} + \widehat{\beta}x_i)\right)^2 .$$

Hvis man i stedet indsætter i (12.4) og reducerer får man et alternativt udtryk for residualkvadratsummen:

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 - \widehat{\beta}^2 \sum_{i=1}^{n}(x_i - \overline{x})^2$$

eller kort

$$SS_y - \widehat{\beta}^2 SS_x \;=\; SS_y - \frac{SP_{xy}^2}{SS_x} \,.$$

Antallet af frihedsgrader er $n-2$ fordi der er n observationer og der er estimeret 2 middelværdiparametre. Skønnet over variansen σ^2 er derfor

$$\begin{aligned} s_{02}^2 &= \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - (\widehat{\alpha} + \widehat{\beta}x_i)\right)^2 \qquad (12.8)\\ &= \frac{1}{n-2}\left(\sum_{i=1}^{n}(y_i - \overline{y})^2 - \widehat{\beta}^2\sum_{i=1}^{n}(x_i - \overline{x})^2\right)\\ &= \frac{1}{n-2}\left(SS_y - \widehat{\beta}^2 SS_x\right) \,. \end{aligned}$$

Parameterskønnenes middelfejl

Regressionsanalyse er i udpræget grad et forsøg på at modellere *kvantitative sammenhænge*, og derfor er det som regel ikke tilstrækkeligt blot at udregne skønnene over parametrene, man skal også skaffe sig en idé om, hvor præcise de er.

Når man *tester hypoteser* foregår det på den måde, at man udregner værdien af en passende valgt teststørrelse, der er indrettet på en måde så den fungerer som et mål for, hvor godt de foreliggende observationer stemmer overens med hypotesen. Derefter bestemmer man den såkaldte testsandsynlighed, der hævdes at være sandsynligheden for at få et sæt observationer der stemmer dårligere overens med hypotesen end de faktiske observationer gør. Når man overhovedet kan tale om en sådan sandsylighed er det takket være den statistiske model, fordi dén fortæller at observationerne kan opfattes som observerede værdier af stokastiske variable der følger en nærmere angivet sandsynlighedsfordeling. Man kan derfor sige, at det vi bruger den statistiske model til er, at sammenligne de faktiske observationer med alle de andre sæt

observationer man også kunne have fået, idet man tager hensyn til, med hvilke sandsynligheder de forekommer.

En anden side af dette at sammenligne med, hvad man ellers kunne have fået, er bestemmelse af *middelfejl på parameterskøn*. Et parameterskøn er jo regnet ud på grundlag af de faktiske observationer, men takket være den statistiske model kan man få svar på spørgsmålet: hvilke andre talværdier af parameterskønnet kunne man også have fået, og med hvilke sandsynligheder. Thi parameterskønnet er en funktion af observationerne, og observationerne kan opfattes som observerede værdier af stokastiske variable, så derfor kan parameterskønnene også opfattes som observerede værdier af nogle stokastiske variable, hvis sandsynlighedsfordeling man i princippet kan finde[12]. Ofte er man endda kun interesseret i at vide, inden for hvilke grænser størstedelen af sandsynlighedsmassen er beliggende, og til det brug udregner man den såkaldte *middelfejl* på parameterskønnet, hvilket er det samme som *parameterskønnets standardafvigelse*. Som en tommelfingerregel gælder nemlig, at intervallet "middelværdien plus/minus to gange standardafvigelsen" afgrænser cirka 95% af sandsynlighedsmassen[13], og i den forstand er middelfejlen et direkte mål for, hvor unøjagtigt parameterskønnet er[14].

Vi skal ikke komme nærmere ind på, *hvordan* man når frem til formeludtryk for middelfejl, men her er nogle resultater for den lineære regressionsmodel:

1. Middelfejlen på $\widehat{\beta}$ er $\sqrt{\frac{1}{SS_x}\,\sigma^2}$.

2. (a) Middelfejlen på $\widehat{\alpha}$ er $\sqrt{\left(\frac{1}{n}+\frac{\bar{x}^2}{SS_x}\right)\sigma^2}$.

   (b) Parameterskønnene $\widehat{\alpha}$ og $\widehat{\beta}$ er korrelerede[15], med en korrela-

[12] Det er sådan set bare en øvelse i transformation af sandsynlighedsfordelinger.

[13] Det er især rigtigt hvis parameterskønnet er normalfordelt.

[14] Eksempel: Hvis man har udregnet $\widehat{\beta}$ til 1.534 og middelfejlen på $\widehat{\beta}$ til 0.3, så véd man, at ca. 95% af alle de andre $\widehat{\beta}$-værdier man også kunne have fået ligger i et interval af længde 1.2 (nemlig intervallet $\beta \pm 2 \times 0.3$), og deraf bør man bl.a. drage den konsekvens at $\widehat{\beta}$ *ikke* skal angives med tre decimaler, men kun med én.

[15] Man bemærker at $\widehat{\alpha}$ og $\widehat{\beta}$ er korrelerede med negativ korrelation. Det kommer af at de to estimater er beregnet ud fra det samme sæt observationer, og det betyder, at hvis observationerne nu f.eks. tilfældigvis er sådan at $\widehat{\beta}$ ligger i den øverste ende af sin fordeling, så vil $\widehat{\alpha}$ med stor sandsynlighed ligge i den nederste ende af *sin* fordeling.

tion som er $-\left(1+\frac{SS_x}{n\overline{x}^2}\right)^{-\frac{1}{2}}$.

3. (a) Middelfejlen på $\widehat{\alpha}+\widehat{\beta}\overline{x}$ er $\sqrt{\frac{1}{n}\sigma^2}$.

   (b) Parameterskønnene $\widehat{\alpha}+\widehat{\beta}\overline{x}$ og $\widehat{\beta}$ er *ukorrelerede*.

Disse udtryk er de teoretiske middelfejl, hvori indgår den teoretiske varians σ^2 på Y. Da vi ikke kender parameteren σ^2 må vi i stedet indsætte et skøn over den, f.eks. s_{02}^2, og derved få de estimerede middelfejl.

Af udtrykket for middelfejlen på $\widehat{\beta}$ ses, at det er en fordel at x-værdierne ligger spredt over et stort interval, for så bliver SS_x stor og middelfejlen derved lille.

À propos middelfejl kan det være værd at nævne, at middelfejlen på et skøn s^2 over variansparameteren σ^2 i en normalfordelingsmodel er lig $\sigma^2\sqrt{2/f}$ hvor f er antallet af frihedsgrader for s^2. Deraf ses hvordan variansskønnet bliver bedre jo flere frihedsgrader det har.

En anden formulering af modellen

I formuleringen af den lineære regressionsmodel (12.2) er der tale om et antal "uspecificerede" talpar (x,y). Ofte er det sådan at der er flere talpar med det samme x, fordi man har foretaget flere målinger af y for hvert x (se f.eks. Eksempel 12.1). Det gør ikke spor; men undertiden er det hensigtsmæssigt at notationen kan indfange dette forhold, bl.a. når man vil lave regneopskrifter der er overkommelige at benytte med "håndkraft". Vi indfører derfor en anden formulering af den lineære regressionsmodel.

I stedet for (12.1) formulerer vi situationen generelt som

baggrundsvariabel	observationer			
x_1	y_{11}	y_{12}	$\cdots$	y_{1n_1}
x_2	y_{21}	y_{22}	$\cdots$	y_{2n_2}
x_3	y_{31}	y_{32}	$\cdots$	y_{3n_3}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
x_k	y_{k1}	y_{k2}	$\cdots$	y_{kn_k}

hvor det nu er sådan at værdierne $x_1, x_2, \ldots, x_k$ af baggrundsvariablen er *forskellige*. Der er altså k forskellige x-værdier, og hørende til den

i-te x-værdi er der de n_i observationer $y_{i1}, y_{i2}, \ldots, y_{in_i}$. Det samlede antal observationer er $n = n_1 + n_2 + \ldots + n_k$.

Regressionsmodellen (12.2) lyder i denne notation

$$Y_{ij} \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2) . \tag{12.9}$$

De tidligere indførte hjælpestørrelser SP_{xy}, SS_x og SS_y (side 271) er i den nye notation

$$\begin{aligned} SP_{xy} &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_i - \overline{x})(y_{ij} - \overline{y}) \\ &= \sum_{i=1}^{k} n_i(x_i - \overline{x})(\overline{y}_i - \overline{y}) \\ &= \sum_{i=1}^{k} n_i x_i \overline{y}_i - \frac{1}{n}\left(\sum_{i=1}^{k} n_i x_i\right)\left(\sum_{i=1}^{k} n_i \overline{y}_i\right) \end{aligned} \tag{12.10}$$

i stedet for (12.5),

$$\begin{aligned} SS_x &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_i - \overline{x})^2 \\ &= \sum_{i=1}^{k} n_i(x_i - \overline{x})^2 \\ &= \sum_{i=1}^{k} n_i x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{k} n_i x_i\right)^2 \end{aligned} \tag{12.11}$$

i stedet for (12.6), og

$$\begin{aligned} SS_y &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2 \\ &= \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{1}{n}\left(\sum_{i=1}^{k} n_i \overline{y}_i\right)^2 \end{aligned}$$

i stedet for (12.7). Her er $\overline{y}_i = \dfrac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$ gennemsnittet af y-erne hørende til x_i, $\overline{y} = \dfrac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} = \dfrac{1}{n}\sum_{i=1}^{k} n_i \overline{y}_i$ er totalgennemsnittet af y-erne,

og $\overline{x} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_i = \frac{1}{n}\sum_{i=1}^{k} n_i x_i$ er gennemsnittet af x-erne (vægtet med n_i-erne).

Parameterestimaterne er stadig

$$\begin{aligned} \widehat{\beta} &= \frac{SP_{xy}}{SS_x}\,, \\ \widehat{\alpha} &= \overline{y} - \widehat{\beta}\overline{x}\,, \text{ og} \\ s_{02}^2 &= \frac{1}{n-2}\left(SS_y - \widehat{\beta}^2 SS_x\right)\,. \end{aligned}$$

Eksempel 12.3. ***Kvælning af hunde, fortsat***

Vi vil antage at hypoxantinkoncentrationen kan beskrives ved en lineær regressionsmodel med hypoxi-varigheden som uafhængig variabel[16].

Vi lader $x_1, x_2, \ldots, x_4$ betegne de fire tidspunkter 0, 6, 12 og 18 min, *og vi lader y_{ij} betegne den j-te koncentrationsværdi til tid x_i. Med de indførte betegnelser er den tidligere foreslåede statistiske model for talmaterialet den lineære regressionsmodel* (12.9).

Vi vil udregne værdierne af estimaterne $\widehat{\alpha}$, $\widehat{\beta}$ og s_{02}^2 over modellens parametre. Man kan selvfølgelig overlade regnearbejdet til en datamat, men det er på den anden side ikke uoverkommeligt at gøre det med håndkraft. Indledningsvis udregnes forskellige hjælpestørrelser mm., se Tabel 12.3.

Heraf fås den estimerede hældningskoefficient til (jf. (12.10) *og* (12.11)*)*

$$\begin{aligned} \widehat{\beta} &= \frac{SP_{xy}}{SS_x} \\ &= \frac{\sum_{i=1}^{k} n_i x_i \overline{y}_i - \frac{1}{n}\left(\sum_{i=1}^{k} n_i x_i\right)\left(\sum_{i=1}^{k} n_i \overline{y}_i\right)}{\sum_{i=1}^{k} n_i x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{k} n_i x_i\right)^2} \end{aligned}$$

[16]Denne antagelse vil blive undersøgt nærmere i en senere fortsættelse af eksemplet, se Eksempel 12.4.

Tabel 12.3: Eksempel 12.3: beregningsskema.
x-værdierne er varighed i min, y-værdierne er koncentration i μmol/l.

i	n_i	x_i	$\overline{y}_i$	$n_i x_i$	$n_i \overline{y}_i$	$n_i x_i \overline{y}_i$	$n_i x_i^2$	$\sum_{j=1}^{n_i} y_{ij}^2$
1	7	0	1.46	0	10.2	0.0	0	22.50
2	6	6	5.50	36	33.0	198.0	216	196.44
3	5	12	7.48	60	37.4	448.8	720	310.26
4	7	18	12.81	126	89.7	1614.6	2268	1197.67
sum	25			222	170.3	2261.4	3204	1726.87

$$= \frac{2261.4 - \dfrac{222 \times 170.3}{25}}{3204 - \dfrac{222^2}{25}} \,\mu\text{mol}\,\text{l}^{-1}\,\text{min}^{-1}$$

$$= \frac{749.14}{1232.64} \,\mu\text{mol}\,\text{l}^{-1}\,\text{min}^{-1}$$

$$= 0.61 \,\mu\text{mol}\,\text{l}^{-1}\,\text{min}^{-1} ,$$

og det estimerede skæringspunkt med ordinataksen til

$$\begin{aligned} \hat{\alpha} &= \overline{y} - \hat{\beta}\,\overline{x} \\ &= \frac{170.3 \,\mu\text{mol}\,\text{l}^{-1}}{25} - \frac{0.61 \,\mu\text{mol}\,\text{l}^{-1}\,\text{min}^{-1} \times 222 \,\text{min}}{25} \\ &= 1.4 \,\mu\text{mol}\,\text{l}^{-1} . \end{aligned}$$

Skønnet over variansparameteren er

$$s_{02}^2 = \frac{1}{n-2}\left(SS_y - \hat{\beta}^2 SS_x\right) .$$

Her er

$$\begin{aligned} SS_y &= \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{1}{n}\left(\sum_{i=1}^{k} n_i \overline{y}_i\right)^2 \\ &= (1726.87 - 170.3^2/25)\,\mu\text{mol}^2\,\text{l}^{-2} \\ &= 566.79 \,\mu\text{mol}^2\,\text{l}^{-2} , \end{aligned}$$

og[17]

$$\begin{aligned}\widehat{\beta}^2 SS_x &= \frac{SP_{xy}^2}{SS_x} \\ &= \frac{749.14^2}{1232.64}\,\mu\text{mol}^2\,\text{l}^{-2} \\ &= 455.29\,\mu\text{mol}^2\,\text{l}^{-2}\,,\end{aligned}$$

så

$$\begin{aligned}s_{02}^2 &= \frac{1}{23}(566.79 - 455.29)\mu\text{mol}^2\,\text{l}^{-2} \\ &= 4.85\,\mu\text{mol}^2\,\text{l}^{-2}\,,\end{aligned}$$

svarende til en estimeret standardafvigelse på 2.2 μmol/l.

Middelfejlen på $\widehat{\beta}$ *er (jf. side 273)*

$$\begin{aligned}\sqrt{\frac{s_{02}^2}{SS_x}} &= \sqrt{\frac{4.85}{1232.64}}\,\mu\text{mol}\,\text{l}^{-1}\,\text{min}^{-1} \\ &= 0.06\,\mu\text{mol}\,\text{l}^{-1}\,\text{min}^{-1}\,,\end{aligned}$$

og middelfejlen på $\widehat{\alpha}$ *er*

$$\begin{aligned}\sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{SS_x}\right) s_{02}^2} &= \sqrt{\left(\frac{1}{25} + \frac{(222/25)^2}{1232.64}\right) 4.85}\,\mu\text{mol}\,\text{l}^{-1} \\ &= 0.7\,\mu\text{mol}\,\text{l}^{-1}\,.\end{aligned}$$

Størrelsen af de to middelfejl viser, at det er passende at angive $\widehat{\beta}$ *med to og* $\widehat{\alpha}$ *med én decimal, så vi må konkludere at den* estimerede regressionslinie *er*

$$y = 1.4\,\mu\text{mol}\,\text{l}^{-1} + 0.61\,\mu\text{mol}\,\text{l}^{-1}\,\text{min}^{-1} \times x\,.$$

□

[17] Bemærk at det ikke er smart at udregne $\widehat{\beta}^2 SS_x$ som $0.61^2 \times 1232.64$. Derved ville afrundingsfejlen i 0.61 nemlig blive ganget med 1232.64, og man ville få resultatet 458.67.

Modelkontrol

Ved lineær regressionsanalyse er den første og vigtigste form for modelkontrol den uhyre simple: at lave en tegning. I et koordinatsystem afsætter man punkterne (x_i, y_i) og man indtegner den *estimerede regressionslinie*[18] og ser efter om det ser ud til at punkterne fordeler sig passende tilfældigt omkring linien[19]. En tegning kan som regel også afsløre hvad der i givet fald måtte være galt med den lineære regressionsmodel.

Tit kan man også foretage et numerisk test for om den lineære regressionsmodel er brugbar. Det foregår ved at man indlejrer regressionsmodellen i en større model, og i denne større model tester man så regressionsmodellen som en statistisk hypotese på helt sædvanlig vis. Hvis det skal kunne lade sig gøre, skal det være sådan at der er flere y-er til det samme x, for man bærer sig ad på en måde at man inddeler y-erne i *grupper*, hvor en gruppe defineres til at bestå af y-er med samme x, og som "større model" benytter man en ensidet variansanalyse-model. Vi benytter den notation der er indført på side 274ff.

Regressionsmodellen

$$Y_{ij} \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2) \tag{12.12}$$

indlejres i en større model, nemlig i den ensidede variansanalyse-model med k grupper svarende til de k niveauer af x:

$$Y_{ij} \sim \mathcal{N}(\mu_i, \sigma^2), \tag{12.13}$$

se f.eks. også Resumé 10. Vi benytter så (12.13) som grundmodel og tester hypotesen (12.12) i forhold hertil. Vi tester altså den hypotese at middelværdierne $\mu_1, \mu_2, \ldots, \mu_k$ ikke er k vilkårlige tal, men at de "hænger sammen" på den måde at der findes to tal α og β således at $\mu_i = \alpha + \beta x_i$ for alle i. Kort sagt testes hypotesen

$$H_2 : \mu_i = \alpha + \beta x_i .$$

Teststørrelsen for at teste H_2 er i princippet en kvotient Q mellem to likelihoodfunktionsværdier, men på samme måde som i Kapitel 11 kan

[18] ikke en "øjemålslinie" ...

[19] Undertiden ligger punktsværmen et helt andet sted end regressionslinien. Så har man enten regnet eller tegnet galt.

Q omskrives til en kvotient F mellem to s^2-størrelser, hvor den i nævneren er variansskønnet i grundmodellen og den i tælleren er "hypotesens variation i forhold til grundmodellen".

Nævneren er variationen inden for grupperne, givet ved den gammelkendte s_0^2 fra (11.3) på side 206:

$$s_0^2 = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_i)^2 .$$

For at bestemme *tælleren* skal vi spalte kvadratsummen

$$(n-2)s_{02}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\overline{y})^2 - \widehat{\beta}^2\sum_{i=1}^{k}n_i(x_i-\overline{x})^2$$

hørende til s_{02}^2 (jf. (12.8)). Fra (11.8) på side 214 vides at

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\overline{y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_i)^2 + \sum_{i=1}^{k}n_i(\overline{y}_i-\overline{y})^2 ,$$

så derfor er

$$(n-2)s_{02}^2 = \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_i)^2}_{(n-k)s_0^2} + \underbrace{\sum_{i=1}^{k}n_i(\overline{y}_i-\overline{y})^2 - \widehat{\beta}^2\sum_{i=1}^{k}n_i(x_i-\overline{x})^2}_{(k-2)s_2^2} .$$

Hermed har vi fået spaltet kvadratsummen hørende til s_{02}^2 (med $n-2$ frihedsgrader) op i en sum af s_0^2-kvadratsummen (med $n-k$ frihedsgrader) og noget mere, som så må være den kvadratsum der skal benyttes til tælleren i F-størrelsen. Sidstnævnte kvadratsum må have $(n-2)-(n-k)=k-2$ frihedsgrader og benævnes derfor $(k-2)s_2^2$, hvor s_2^2 er et variansskøn der beskriver *gruppegennemsnittenes variation omkring regressionslinien.*

Teststørrelsen for linearitetshypotesen H_2 er altså

$$F = \frac{s_2^2}{s_0^2} , \tag{12.14}$$

hvor

$$s_2^2 = \frac{1}{k-2}\left(\sum_{i=1}^{k}n_i(\overline{y}_i-\overline{y})^2 - \widehat{\beta}^2\sum_{i=1}^{k}n_i(x_i-\overline{x})^2\right) \tag{12.15}$$

og hvor s_0^2 er givet ovenfor. Store værdier af F er signifikante, og hvis H_2 er rigtig vil F følge F-fordelingen med frihedsgrader $k-2$ og $n-k$, således at testsandsynligheden ε er givet som

$$\varepsilon \;=\; \mathrm{P}\left(F_{k-2,n-k} > F_{\mathrm{obs}}\right)$$

der bestemmes ved hjælp af en tabel over F-fordelingen.

Dette test bør *godkende* linearitetshypotesen H_2 for at vi kan gå videre med den lineære regressionsmodel.

Eksempel 12.4. ***Kvælning af hunde, fortsat***

Vi vil undersøge, om det kan antages at hypoxantin-koncentrationen afhænger lineært af hypoxiens varighed. Da vi er i en situation hvor der er en del y-er til hvert x, er det muligt at udføre det numeriske test for modellen.

Vi har tidligere (i Eksempel 12.3) bestemt de talværdier der i givet fald er de bedste skøn over parametrene og derved fået den estimerede regressionslinie til

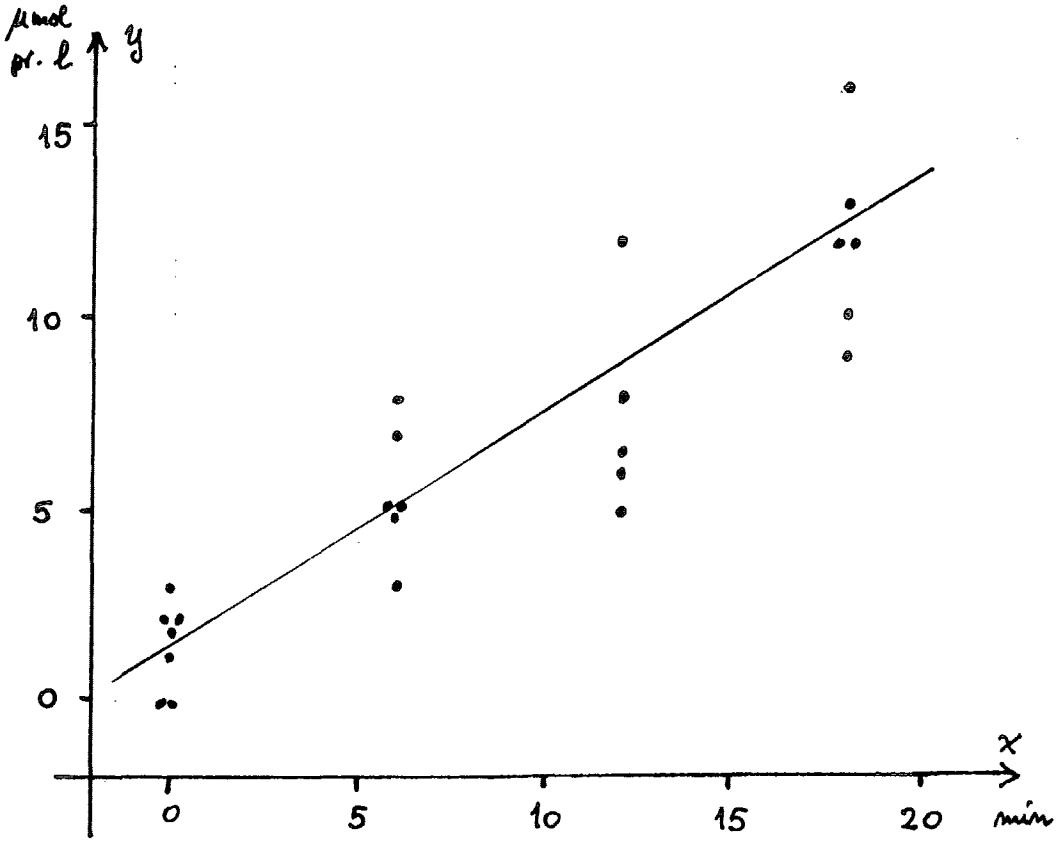
$$y \;=\; 1.4\,\mu\mathrm{mol\,l^{-1}} + 0.61\,\mu\mathrm{mol\,l^{-1}\,min^{-1}} \times x\ .$$

I Figur 12.1 er indtegnet dels sammenhørende værdier af varighed og koncentration, dels den estimerede regressionslinie. Efter tegningen at dømme er den lineære regressionsmodel ikke helt hen i vejret. For at bestyrke troen på modellen vil vi udføre det numeriske test for den lineære model.

Som en midlertidig grundmodel vil vi benytte en ensidet variansanalyse-model baseret på de fire grupper bestemt af x-erne. Indledningsvis udregnes forskellige hjælpestørrelser mm., se Tabel 12.4. Det fremgår bl.a. at den kvadratsum der beskriver variationen mellem grupper *er 101.32 med 21 frihedsgrader, og nævneren i teststørrelsen* (12.14) *for H_2 er dermed $s_0^2 = 4.82\mu\mathrm{mol}^2\mathrm{l}^{-2}$. Tælleren i teststørrelsen er givet ved* (12.15). *I Eksempel 12.3 fandt vi (side 278) at*

$$\begin{aligned}
\widehat{\beta}^2 \sum_{i=1}^{k} n_i(x_i - \overline{x})^2 \;&=\; \widehat{\beta}^2 SS_x \\
&=\; 455.29\ , \qquad \textit{og} \\
\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2 \;&=\; 566.79\ .
\end{aligned}$$

Figur 12.1: Eksempel 12.4: Sammenhørende værdier af hypoxantinkoncentration og hypoxivarighed, samt den estimerede regressionslinie.



Tabel 12.4: Eksempel 12.4: Nogle hjælpestørrelser.

i	n_i	$\sum_{j=1}^{n_i} y_{ij}$	$\overline{y}_i$	f_i	$\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_i)^2$	s_i^2
1	7	10.2	1.46	6	7.64	1.27
2	6	33.0	5.50	5	14.94	2.99
3	5	37.4	7.48	4	30.51	7.63
4	7	89.7	12.81	6	48.23	8.04
sum	25	170.3		21	101.32	
gennemsnit			6.81			4.82

Vi har brug for værdien af $\sum_{i=1}^{k} n_i(\overline{y}_i - \overline{y})^2$, *og den kan man ifølge* (11.8) *på side 214 udregne som*

$$\begin{aligned}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\overline{y})^2 - \sum_{i=1}^{k} n_i(y_{ij}-\overline{y}_i)^2 &= 566.79 - 101.32 \\ &= 465.47\end{aligned}$$

med $24 - 21 = 3$ *frihedsgrader. Tællerkvadratsummen, dvs. kvadratsummen der beskriver gruppegennemsnittenes variation omkring regressionslinien, er derfor* $465.47 - 455.29 = 10.18$ *med* $k - 2 = 2$ *frihedsgrader, således at det tilsvarende variansskøn er* $s_2^2 = 5.09\ \mu\text{mol}^2\,\text{l}^{-2}$.

F-teststørrelsen der måler gennemsnittenes variation omkring linien i forhold til variationen inden for grupper er dermed

$$\begin{aligned}F_{\text{obs}} &= \frac{s_2^2}{s_0^2} \\ &= \frac{5.09}{4.82} \\ &= 1.06\end{aligned}$$

der skal sammenlignes med F-fordelingen med 2 og 21 frihedsgrader, og i denne fordeling er der mere end 30% sandsynlighed

Tabel 12.5: Eksempel 12.4: Variansanalyseksema vedr. test af linearitetshypotesen.

f står for antal frihedsgrader, SS står for sum af kvadratiske afvigelser, $s^2 = SS/f$.

variation	f	SS	s^2	test
inden for grupper	21	101.32	4.82	
gennemsnittenes var. omkring regr.linien	2	10.18	5.09	5.09/4.82=1.06
samlet variation omkring regr.linien	23	111.50	4.85	

for at få en værdi som er større end den observerede, der altså på ingen måder er signifikant. Vi har således fået bekræftet linearitetshypotesen.

Traditionelt opsummerer man udregninger og testresultater i et variansanalyseskema, se Tabel 12.5.

•

Variansanalysemodellen såvel som den lineære regressionsmodel forudsætter at der er varianshomogenitet, så det kan man jo også teste. Vi indsætter s^2-værdierne fra Tabel 12.4 i Bartlett's teststørrelse (se f.eks. Resumé 9) og får

$$\begin{aligned} B &= -\left(6 \ln \frac{1.27}{4.82} + 5 \ln \frac{2.99}{4.82} + 4 \ln \frac{7.63}{4.82} + 6 \ln \frac{8.04}{4.82}\right) \\ &= 5.5 \,, \end{aligned}$$

der skal sammenlignes med χ^2-fordelingen med $k - 1 = 3$ frihedsgrader. Tabelopslag viser at der er over 10% chance for at få en større B-værdi end værdien 5.5, som derfor ikke er signifikant stor. Med andre ord, vi kan opretholde antagelsen om varianshomogenitet.

•

Alt i alt er der således ikke noget der taler i mod at vi beskriver hypoxi-dataene med en lineær regressionsmodel med hypoxivarighed som uafhængig variabel og hypoxantinkoncentration som afhængig variabel. □

Test af hypoteser om liniens parametre

Man kan naturligvis teste hypoteser om regressionsliniens parametre. Fremgangsmåden er den samme som altid: først estimeres parametrene under hypotesen, dernæst udregnes kvotienten Q mellem de to maksimale likelihoodfunktionsværdier, og endelig bestemmes sandsynligheden for at få et værre sæt observationer, dvs. et sæt observationer der giver et mindre Q. Som ved alle andre tests af middelværdihypoteser i normalfordelingen kan Q omskrives til en F-størrelse, der er mere praktisk at have med at gøre, og da de hypoteser der er tale om, er hypoteser om en enkelt parameter, kan F-teststørrelsen yderligere omdannes til en t-teststørrelse der måske er mere forståelig.

Vi skal ikke her komme ind på de nærmere detaljer, men blot forklare hvordan teststørrelserne kommer til at se ud i disse specielle tilfælde.

Hypoteser om hældningskoefficienten

Hvis man vil teste hypotesen

$$H_3 : \beta = 0$$

om at der ikke er nogen signifikant lineær sammenhæng mellem y og x, så bliver F-teststørrelsen

$$F = \frac{s_3^2}{s_{02}^2}$$

hvor s_{02}^2 fra (12.8) er det bedste variansskøn under den aktuelle model, og hvor s_3^2 er $\hat{\beta}^2 SS_x$. Store værdier af F er signifikante.

Man kan også skrive F-størrelsen som t^2, hvor

$$t = \frac{\hat{\beta}}{\sqrt{s_{02}^2/SS_x}}$$

er skønnet $\widehat{\beta}$ over β divideret med skønnet over standardafvigelsen på $\widehat{\beta}$ (eller med skønnet over middelfejlen på $\widehat{\beta}$), jf. side 273. t-størrelsen måler således hvor langt $\widehat{\beta}$ ligger fra den formodede værdi 0, når man benytter middelfejlen som målestok. Store værdier af $|t|$ er signifikante.

Man kan bevise at under H_3 vil t være t-fordelt med det antal frihedsgrader som s_{02}^2 har, dvs. med $n-2$ frihedsgrader. Det betyder at testsandsynligheden kan findes ved hjælp af tabeller over t-fordelingen som[20]

$$\varepsilon \;=\; \mathrm{P}\left(|t_{n-2}| > |t_{\mathrm{obs}}|\right) \;.$$

Hvis hypotesen H_3 kan godkendes, skal man udregne et revideret skøn over α og et forbedret skøn over variansen σ^2. H_3 betyder jo, at den "forklarende" variabel x ikke er nødvendig, men at alle Y-er har samme middelværdi α, dvs. der er tale om et "enstikprøveproblem". Hvis H_3 er rigtig er skønnet over α derfor totalgennemsnittet $\overline{y}$ og skønnet over σ^2 er

$$s_{03}^2 \;=\; \frac{1}{n-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\overline{y})^2 \;.$$

Hypoteser om skæringen med y-aksen

Undertiden følger det af den faglige problemstilling at linien *skal* gå gennem $(0,0)$, dvs. at $\alpha = 0$, i andre situationer kan man være interesseret i at teste hypoteser om α blot for at nå til en så simpel beskrivelse af data som muligt. Hvis man ønsker at teste hypotesen

$$H_4 \;:\; \alpha = 0$$

om at linien går gennem $(0,0)$, kan det gøres med F-teststørrelsen

$$F \;=\; \frac{s_4^2}{s_{02}^2} \;,$$

hvor tælleren er "kvadratsummen"

$$\frac{n\,SS_x\,\widehat{\alpha}^2}{SS_x + n\overline{x}^2}$$

[20] Hvis man vil benytte F som teststørrelse er $\varepsilon = \mathrm{P}(F_{1,n-2} > F_{\mathrm{obs}})$.

divideret med sit frihedsgradsantal 1 og nævneren er variansskønnet under linearitetshypotesen[21]. Store værdier af F er signifikante.

Det er dog lettere at udnytte at $F = t^2$ hvor

$$t = \frac{\hat{\alpha}}{\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{SS_x}\right) s_{02}^2}}$$

er forholdet mellem skønnet $\hat{\alpha}$ over α og skønnet over middelfejlen på $\hat{\alpha}$. Store værdier af $|t|$ er signifikante.

Man kan bevise at under H_4 vil t følge t-fordelingen med samme antal frihedsgrader som variansskønnet i nævneren, dvs. $n-2$ frihedsgrader. Det betyder at testsandsynligheden kan findes ved hjælp af tabeller over t-fordelingen som

$$\varepsilon = \mathrm{P}\left(|t_{n-2}| > |t_{\mathrm{obs}}|\right) .$$

Hvis hypotesen H_4 kan godkendes skal man udregne et revideret skøn over hældningen β og et forbedret skøn over σ^2. Det nye skøn over β bliver

$$\hat{\hat{\beta}} = \frac{\sum_{i=1}^{k} n_i x_i \bar{y}_i}{\sum_{i=1}^{k} n_i x_i^2} ,$$

og skønnet over σ^2 bliver

$$\begin{aligned} s_{04}^2 &= \frac{1}{n-1} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \hat{\hat{\beta}} x_i\right)^2 \\ &= \frac{1}{n-1} \left(\sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}^2 - \hat{\hat{\beta}}^2 \sum_{i=1}^{k} n_i x_i^2 \right) . \end{aligned}$$

Nogle bemærkninger om kvadratsummer

I kapitlets løb er der efterhånden kommet så mange kvadratafvigelsessummer (med tilhørende frihedsgrader) i spil, at det måske kan knibe

[21] Man tester såvel H_3 som H_4 i forhold til linearitetsmodellen (12.9).

at bevare overblikket over dem. Ydermere er der to forskellige diskussioner der føres, idet der dels er spørgsmålet om den statistiske (og faglige) betydning af kvadratsummerne, dels spørgsmålet om hvordan man lettest udregner summerne.

Hvis vi tager betydningsspørgsmålet først så skal man have in mente, at det spil der foregår går ud på at spalte den totale variation op i noget der forklares af diverse middelværdiparametre og noget andet, resten, som kaldes tilfældigt og som beskrives af den såkaldte residualkvadratsum. Når man tester en statistisk hypotese, tester man om man kan klare sig med færre middelværdiparametre end først antaget, og det gøres på den måde at man ved hjælp af F-teststørrelsen[22] ser efter om den tilsvarende forøgelse af residualkvadratsummen er lille i sammenligning med den allerede benyttede/accepterede residualkvadratsum.

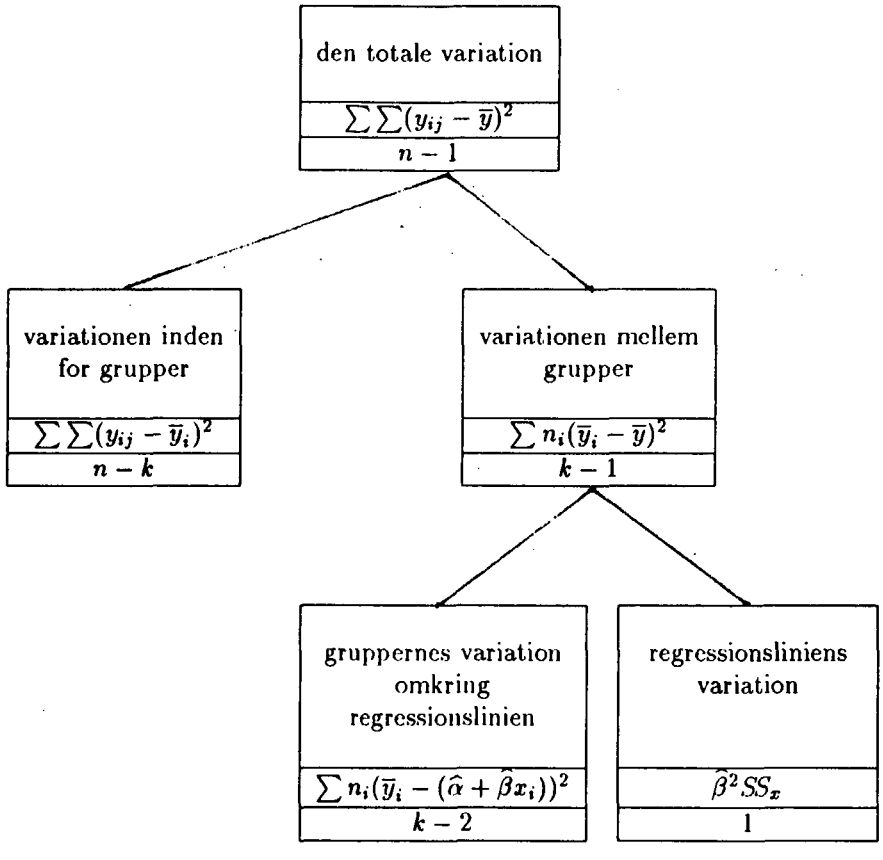
Figur 12.2 giver en oversigt over indeværende kapitels kvadratsummer og deres opspaltning. I første omgang spaltes *den totale variation* op i *variationen mellem grupper*, som beskrives af k middelværdiparameter, og *variationen inden for grupper* som har rollen af residualkvadratsum. Dernæst spaltes variationen mellem grupper op som vist i figuren, og testet for linearitetshypotesen H_2 består i at vurdere om *gruppernes variation omkring linien* er så tilpas lille at det er rimeligt at inkludere den i residualvariationen. I givet fald får vi en residualvariation som har $(n-k)+(k-2) = k-2$ frihedsgrader og som består af variationen inden for grupper plus gruppernes variation omkring linien; denne variation kan kaldes den *samlede variation omkring regressionslinien*, og det er den der danner grundlag for variansskønnet s_{02}^2. Når man tester hypotesen H_3 om at hældningskoefficienten er 0, så består det i at vurdere om *regressionsliniens variation* er så tilpas lille at også dén kan inkluderes i residualvariationen.

Med hensyn til beregningsmetoder kan det være noget af en smags sag hvad man finder lettest, men den strategi der blev benyttet i eksemplet var følgende:

1. Beregn den totale kvadratsum $SS_y = \sum\sum(y_{ij} - \overline{y})^2$.
2. Beregn kvadratsummen "inden for grupper" $\sum(\sum(y_{ij} - \overline{y}_i)^2)$ (der også skal benyttes hvis man vil lave Bartlett's test).
3. Så er "mellem grupper"-kvadratsummen lig med den totale kvadratsum minus kvadratsummen "inden for grupper".

[22] eller en t-teststørrelse hvis man kun tester én parameter bort

Figur 12.2: Oversigt over nogle af de kvadratsummer med tilhørende frihedsgradsantal som optræder i dette kapitel.



4. I forbindelse med udregning af $\widehat{\beta}$ udregnes kvadratsummen "regressionsliniens variation", dvs. $\widehat{\beta}^2 SS_x = SP_{xy}^2/SS_x$.

5. Så er kvadratsummen "gruppernes variation omkring linien" lig med "variationen mellem grupper" minus "regressionsliniens variation".

Dette gælder for *kvadratsummer* og for *frihedsgrader*, men ikke for variansskøn. Variansskøn s^2 udregnes altid som den relevante kvadratsum divideret med det tilhørende frihedsgradsantal.

Resumé 14. Simpel lineær regressionsanalyse

Situation: Der foreligger n sammenhørende par (x_i, y_i) af en observation y og en baggrundsvariabel x; skematisk ser det sådan ud:

observation	baggrundsvariabel
y_1	x_1
y_2	x_2
$\vdots$	$\vdots$
y_n	x_n

Model: Det antages at x_i-erne er givne konstanter og at y_i-erne er observerede værdier af uafhængige stokastiske variable Y_i, således at Y_i er normalfordelt med middelværdi $\alpha + \beta x_i$ og varians σ^2, kort

$$Y_i \sim \mathcal{N}(\alpha + \beta x_i\,,\, \sigma^2)\,,$$

hvor α, β og σ^2 er ukendte parametre. Herved beskriver de to middelværdiparametre α og β den systematiske variation, nemlig ys lineære afhængighed af x, og variansparametreren σ^2 (samt normalfordelingen) beskriver den tilfældige variation omkring regressionslinien; den tilfældige variation antages at være den samme for alle x.

Estimation: Udregn hjælpestørrelserne

$$\begin{aligned}
\overline{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i\,,\\
\overline{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i\,,\\
SP_{xy} &= \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})\,,\\
SS_x &= \sum_{i=1}^{n}(x_i - \overline{x})^2\,,\\
SS_y &= \sum_{i=1}^{n}(y_i - \overline{y})^2\,.
\end{aligned}$$

Så er skønnene over α og β givet ved

$$\widehat{\beta} = \frac{SP_{xy}}{SS_x}\,,$$

$$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x} .$$

Variansparameteren σ^2 estimeres ved residualkvadratsummen divideret med antallet af frihedsgrader, nemlig

$$s_{02}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - (\hat{\alpha} + \hat{\beta}x_i)\right)^2$$

$$= \frac{1}{n-2}\left(SS_y - \hat{\beta}^2 SS_x\right)$$

der har $n-2$ frihedsgrader.

Modelkontrol: Man kan lave en tegning hvor man i et koordinatsystem afsætter sammenhørende værdier af x og y og desuden indtegner den estimerede regressionslinie. Punkterne skal da fordele sig nogenlunde tilfældigt omkring linien.

Hvis der til hver x-værdi er flere y-observationer kan man foretage et numerisk test for antagelsen om linearitet, se Resumé 15.

Hypoteser:

1. Hypotesen $H_3 : \beta = 0$ om at y ikke afhænger signifikant af x testes med t-teststørrelsen

$$t = \frac{\hat{\beta}}{\sqrt{s_{02}^2/SS_x}} .$$

Store værdier af $|t|$ er signifikante.

2. Hypotesen $H_4 : \alpha = 0$ om at regressionslinien går gennem $(0,0)$ testes med t-teststørrelsen

$$t = \frac{\hat{\alpha}}{\sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{SS_x}\right)s_{02}^2}} .$$

Store værdier at $|t|$ er signifikante.

I begge tilfælde bestemmes **testsandsynligheden** ε som sandsynligheden for at få en værdi uden for intervallet med endepunkter $-t_{\text{obs}}$ og t_{obs} i t-fordelingen med $n-2$ frihedsgrader,

$$\varepsilon = 2 \times \mathrm{P}\left(t_{n-2} > |t_{\text{obs}}|\right) .$$

Hvis ε er meget lille, så er **konklusionen** at den pågældende parameter er signifikant forskellig fra 0, dvs. hypotesen forkastes. Hvis ε ikke er meget lille, kan man ikke på det foreliggende grundlag forkaste den pågældende hypotese.

Resumé 15. Test for linearitet i den lineære regressionsmodel

Dette resumé skal læses i sammenhæng med Resumé 14, *Simpel lineær regressionsanalyse*, og handler om det særlige tilfælde hvor der er flere y-værdier til hvert x.

Situation: Til hver af k forskellige værdier af baggrundsvariablen x foreligger et antal observationer y som det fremgår af nedenstående skema:

baggrundsvariabel	observationer
x_1	$y_{11}\ y_{12}\ \cdots\ y_{1n_1}$
x_2	$y_{21}\ y_{22}\ \cdots\ y_{2n_2}$
x_3	$y_{31}\ y_{32}\ \cdots\ y_{3n_3}$
$\vdots$	$\vdots\ \vdots\ \ddots\ \vdots$
x_k	$y_{k1}\ y_{k2}\ \cdots\ y_{kn_k}$

Det totale antal y-observationer er $n = n_1 + n_2 + \ldots + n_k$.

Model: Det antages at $x_1, x_2, \ldots, x_k$ er k forskellige givne konstanter og at y_{ij}-erne er observerede værdier af uafhængige stokastiske variable Y_{ij}, således at Y_{ij} er normalfordelt med middelværdi $\alpha + \beta x_i$ og varians σ^2, kort

$$Y_{ij} \sim \mathcal{N}(\alpha + \beta x_i\,,\, \sigma^2)\,,$$

hvor α, β og σ^2 er ukendte parametre. Herved beskriver de to middelværdiparametre α og β den systematiske variation, nemlig ys lineære afhængighed af x, og variansparametreren σ^2 (samt normalfordelingen) beskriver den tilfældige variation omkring regressionslinien; den tilfældige variation antages at være den samme for alle x.

Estimation: Udregn hjælpestørrelserne[23]

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i\,,$$

$$\overline{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}\,,$$

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}\,,$$

[23]der (med undtagelse af $\overline{y}_i$) er de samme som i Resumé 14.

$$SP_{xy} = \sum_{i=1}^{k} n_i(x_i - \overline{x})(\overline{y}_i - \overline{y}) ,$$

$$SS_x = \sum_{i=1}^{k} n_i(x_i - \overline{x})^2 ,$$

$$SS_y = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \overline{y})^2 .$$

Desuden udregnes $\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)^2$.

Så er skønnene over α og β givet ved

$$\widehat{\beta} = \frac{SP_{xy}}{SS_x} ,$$

$$\widehat{\alpha} = \overline{y} - \widehat{\beta}\overline{x} .$$

Variansparameteren σ^2 estimeres ved residualkvadratsummen divideret med antallet af frihedsgrader, nemlig

$$s_{02}^2 = \frac{1}{n-2} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - (\widehat{\alpha} + \widehat{\beta}x_i)\right)^2$$

$$= \frac{1}{n-2} \left(SS_y - \widehat{\beta}^2 SS_x\right)$$

der har $n - 2$ frihedsgrader.

Modelkontrol: Man ønsker at *teste* linearitetsantagelsen med et numerisk test, og derfor formuleres en ny model, nemlig den ensidede variansanalysemodel (jf. Resumé 10)

$$Y_{ij} \sim \mathcal{N}(\mu_i , \sigma^2)$$

og i forhold hertil testes den oprindelige model som en hypotese. **Teststørrelsen** er

$$F = \frac{s_2^2}{s_0^2}$$

hvor

$$s_0^2 = \frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)^2 \quad \text{og}$$

$$s_2^2 = \frac{1}{k-2}\sum_{i=1}^{k} n_i \left(\overline{y}_i - (\widehat{\alpha} + \widehat{\beta} x_i)\right)^2$$

$$= \frac{1}{k-2}\left(SS_y - \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2 - \widehat{\beta}^2 SS_x\right) .$$

– Store værdier af F er signifikante.

Hvis linearitetshypotesen er rigtig vil F være F-fordelt med $k-2$ og $n-k$ frihedsgrader, så **testsandsynligheden** ε bestemmes som

$$\varepsilon = \mathrm{P}\left(F_{k-2,n-k} > F_{\mathrm{obs}}\right) .$$

Hvis ε ikke er meget lille, så er der ikke nogen signifikant afvigelse fra linearitet, dvs. linearitetsmodellen kan opretholdes.

12.2 Multipel lineær regressionsanalyse

Ofte ønsker man at opbygge en regressionsmodel der inddrager mere end én forklarende variabel. Vi vil derfor nu betragte den situation hvor der for hvert af et antal "individer" foreligger dels en observation y, dels værdier $x_1, x_2, \ldots, x_p$ af p baggrundsvariable: Til individ nr. i hører observationen y_i og værdierne $x_{i1}, x_{i2}, \ldots, x_{ip}$ af baggrundsvariablene. Skematisk ser det sådan ud:

observation	baggrundsvariabel			
y_1	x_{11}	x_{12}	$\ldots$	x_{1p}
y_2	x_{21}	x_{22}	$\ldots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
y_n	x_{n1}	x_{n2}	$\ldots$	x_{np}

Den statistiske model for y-erne indrettes på følgende måde:

- $y_1, y_2, \ldots, y_n$ er observerede værdier af stokastiske variable $Y_1, Y_2, \ldots, Y_n$.
- De stokastiske variable $Y_1, Y_2, \ldots, Y_n$ er uafhængige og normalfordelte med samme varians σ^2.
- x-erne betragtes som faste tal - de er altså ikke observerede værdier af stokastiske variable.
- Middelværdien af Y-erne er en lineær funktion af $x_1, x_2, \ldots, x_p$, således at forstå at der findes $p+1$ parametre $\alpha, \beta_1, \beta_2, \ldots, \beta_p$ så

$$\mathrm{E}\,Y_i = \alpha + \sum_{j=1}^{p} x_{ij}\beta_j$$

for alle i.

Af æstetiske grunde indfører man gerne en ekstra baggrundsvariabel x_0 der er lig med 1 for alle i, og samtidig kalder man α for β_0. Så er nemlig

$$\mathrm{E}\,Y_i = \sum_{j=0}^{p} x_{ij}\beta_j$$

for alle i.

Modellen skrives kortfattet som

$$\mathrm{E}\,Y_i \;=\; \sum_{j=0}^{p} x_{ij}\beta_j$$

eller bedre

$$Y_i \;\sim\; \mathcal{N}(\sum_{j=0}^{p} x_{ij}\beta_j\,,\,\sigma^2)\,. \tag{12.16}$$

Denne model er en såkaldt *multipel lineær regressionsmodel*. Den beskriver y-ernes *systematiske variation* ved hjælp af de $p+1$ ukendte middelværdiparametre $\beta_0, \beta_1, \ldots, \beta_p$ plus de kendte konstanter x_{ij}, og den beskriver den tilfældige variation ved hjælp af normalfordelingen og den ukendte variansparameter σ^2.

Estimation af parametrene

Som altid estimeres parametrene i modellen ved at maksimalisere likelihoodfunktionen, der i det foreliggende tilfælde ser sådan ud:

$$\begin{aligned} & L(\beta_0, \beta_1, \ldots, \beta_p, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\tfrac{1}{2} \frac{\left(y_i - \sum_{j=0}^{p} x_{ij}\beta_j \right)^2}{\sigma^2} \right) \\ &= (2\pi)^{-n/2} (\sigma^2)^{-n/2} \exp\left(-\tfrac{1}{2} \frac{\sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} x_{ij}\beta_j \right)^2}{\sigma^2} \right) . \end{aligned}$$

Heraf ses at de bedste skøn over middelværdiparametrene er dem der minimaliserer kvadratsummen

$$\sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} x_{ij}\beta_j \right)^2 .$$

Denne gang er der ingen vej uden om de lidt mere avancerede matematiske metoder vedrørende minimalisering af funktioner af flere variable[24]. Disse metoder fortæller at minimumspunktet findes som det punkt hvor de $p+1$ partielle afledede (mht. de $p+1$ β-er) er lig 0. Hvis man skriver op hvad det betyder og omskriver en smule, når man frem til $p+1$ ligninger[25] med $p+1$ ubekendte. Den j-te af disse ligninger ser således ud:

$$\sum_{l=0}^{p}\left(\sum_{i=1}^{n} x_{ij}x_{il}\right)\beta_l \;=\; \sum_{i=1}^{n} x_{ij}y_i \;. \qquad (12.17)$$

Det kan godt se lidt overvældende ud, men "i virkeligheden" står der bare

$$a_{j0}\beta_0 + a_{j1}\beta_1 + \ldots + a_{jp}\beta_p \;=\; c_j$$

hvor a-erne og c er de kendte tal

$$a_{jl} \;=\; \sum_{i=1}^{n} x_{ij}x_{il} \;,$$

$$c_j \;=\; \sum_{i=1}^{n} x_{ij}y_i \;,$$

og $\beta_0, \beta_1, \ldots, \beta_p$ er ligningernes ubekendte.

Ved at løse de $p+1$ ligninger (12.17) får man de bedste skøn $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$[26]. Dernæst kan man udregne residualkvadratsummen

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p} x_{ij}\hat{\beta}_j\right)^2$$

og variansskønnet

$$s_0^2 \;=\; \frac{1}{n-(p+1)}\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p} x_{ij}\hat{\beta}_j\right)^2$$

der har $n-(p+1)$ frihedsgrader.

[24] Vi var ude for et tilsvarende problem i den multiplikative Poissonmodel, se side 139.

[25] der i matrix-notation ganske enkelt er $(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{y}$.

[26] Der er præcis én løsning til ligningerne, medmindre man har valgt baggrundsvariablene så uheldigt at nogle af dem indeholder information der allerede er indeholdt i de andre.

Modelkontrol

I tilfældet $p = 1$ der behandledes i Afsnit 12.1 kan man kontrollere sin model ved hjælp af enkle tegninger, hvilket ikke lader sig gøre når $p > 1$. Man må derfor finde på andre metoder. Een ting der er fornuftig at foretage sig er at udregne residualerne

$$e_i = y_i - \sum_{j=0}^{p} x_{ij}\hat{\beta}_j$$

og se hvordan de fordeler sig. Hvis modellen (12.16) er rigtig, er de *teoretiske* residualer

$$y_i - \sum_{j=0}^{p} x_{ij}\beta_j$$

uafhængige $\mathcal{N}(0\,,\,\sigma^2)$-fordelte. Vi kender kun de empiriske residualer $e_1, e_2, \ldots, e_n$, og der gælder at hvis modellen passer så vil de være $\mathcal{N}(0\,,\,\sigma^2)$-fordelte og næsten uafhængige[27]. Man må derfor se efter om residualerne ser ud til at være nogenlunde uafhængige og normalfordelte.

I Afsnit 12.1 omtaltes et *numerisk test* for linearitetshypotesen. Dette test kunne udføres når der var flere y-værdier til hvert enkelt x, således at man kunne indføre nogle grupper og bestemme en variation inden for grupper. Når der er tale om *multipel* regressionsanalyse kan man gøre noget tilsvarende, forudsat at der er flere y-værdier for hvert enkelt sæt $(x_1, x_2, \ldots, x_p)$ af baggrundsvariable. Denne forudsætning er sædvanligvis kun opfyldt hvis man ved planlægningen af forsøget udtrykkelig har taget hensyn til den.

Antallet af baggrundsvariable

Undertiden foreligger der et større sortiment af baggrundsvariable, og en og anden ville måske mene at det var en fordelagtig situation. Men hvis man – i et ekstremt tilfælde – inddrog lige så mange baggrundsvariable som der var observationer, så ville man ganske vist opnå et perfekt fit, men samtidig ville man demonstrere en total mangel på forståelse af den statistiske models rolle. Formålet med statistiske metoder er at skille *det systematiske* og *det tilfældige* og at beskrive begge dele på

[27] jo flere frihedsgrader der er, jo mere uafhængige er de.

simpel vis, og det indebærer blandt andet at man ikke skal inddrage flere baggrundsvariable i den endelig model end højst nødvendigt.

Som regel vil man gerne kunne "forstå" modellerne, så derfor må det være fornuftigt at man så vidt muligt kun indddrager de af baggrundsvariablene der kan formodes at have en betydning. Når man har udpeget de formodet interessante baggrundsvariable, skal man bestemme hvilke af dem der faktisk skal med i modellen, og det er ikke nogen simpel opgave: skal man begynde med én variable (hvilken?) og så inddrage flere og flere indtil modellen giver en tilstrækkelig god beskrivelse, eller skal man begynde med at have alle variable inde i modellen og så fjerne de ikke-signifikante en efter en? Og efter hvilket kriterium skal man tage variable ind og ud af modellen? Vi skal ikke her komme ind på disse spørgsmål, som der kan skrives tykke bøger om (– og det er der også).

Eksempel 12.5. ***Indianere i Peru***

Ændringer i menneskers livsbetingelser kan give sig udslag i fysiologiske ændringer, eksempelvis i ændret blodtryk.

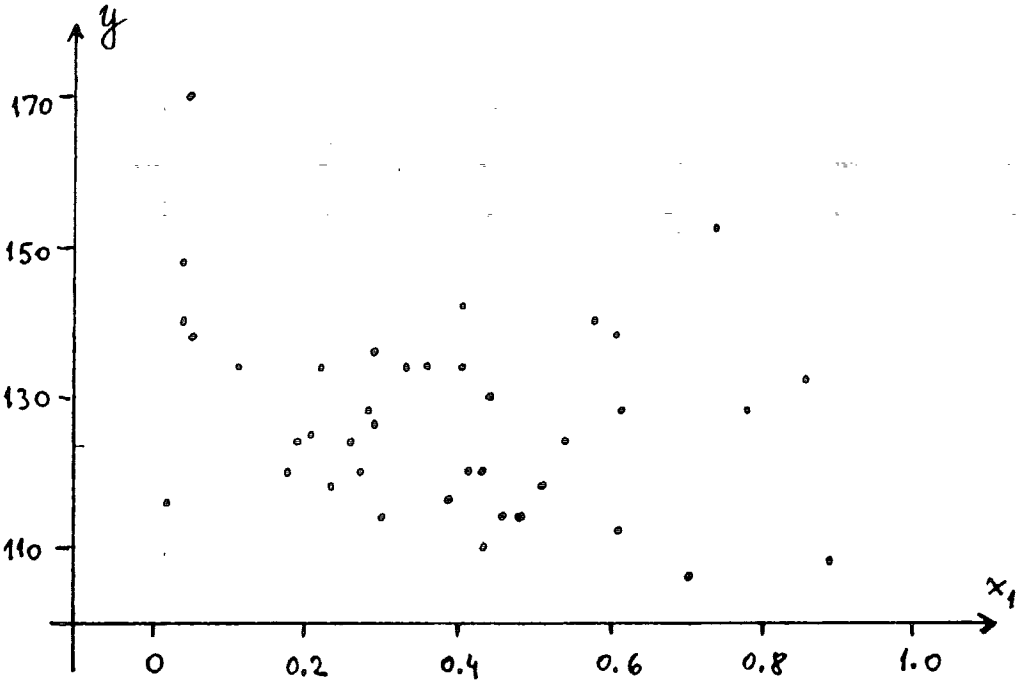
En gruppe antropologer har undersøgt hvordan blodtrykket ændrer sig hos peruvianske indianere der flyttes fra deres oprindelige primitive samfund i de høje Andesbjerge til den såkaldte civilisation, dvs. storbyen, der i øvrigt ligger i langt mindre højde over havets overflade end deres oprindelig bopæl. Antropologerne udvalgte en stikprøve på 39 mænd over 21 år der havde undergået en sådan flytning. På hver af disse måltes blodtrykket (både det systoliske og det diastoliske) samt en række baggrundsvariable, bl.a. alder, antal år siden flytningen, højde, vægt og puls. Som om det ikke kunne være nok har man udregnet endnu en baggrundsvariabel, nemlig "brøkdel af livet levet i de nye omgivelser", dvs. antal år siden flytning divideret med nuværende alder. Man kunne forestille sig at denne baggrundsvariabel ville have stor "forklaringsevne".

Her vil vi ikke se på hele talmaterialet men kun på blodtrykket *(det systoliske) der skal optræde som y-variabel og på de to x-variable* brøkdel af livet i de nye omgivelser *og* vægt. *Disse er angivet i Tabel 12.6.*

Tabel 12.6: Eksempel 12.5: Værdier af y: systolisk blodtryk (mm Hg), x_1: brøkdel af livet i de nye omgivelser, og x_2: vægt (kg).

y	x_1	x_2
170	0.048	71.0
120	0.273	56.5
125	0.208	56.0
148	0.042	61.0
140	0.040	65.0
106	0.704	62.0
120	0.179	53.0
108	0.893	53.0
124	0.194	65.0
134	0.406	57.0
116	0.394	66.5
114	0.303	59.1
130	0.441	64.0
118	0.514	69.5
138	0.057	64.0
134	0.333	56.5
120	0.417	57.0
120	0.432	55.0
114	0.459	57.0
124	0.263	58.0
114	0.474	59.5
136	0.289	61.0
126	0.289	57.0
124	0.538	57.5
128	0.615	74.0
134	0.359	72.0
112	0.610	62.5
128	0.780	68.0
134	0.122	63.4
128	0.286	68.0
140	0.581	69.0
138	0.605	73.0
118	0.233	64.0
110	0.432	65.0
142	0.409	71.0
134	0.222	60.2
116	0.021	55.0
132	0.860	70.0
152	0.741	87.0

Figur 12.3: Eksempel 12.5: Blodtryk y afsat mod brøkdel af livet siden flytning x_1.



Antropologerne mente at x_1 (brøkdel levet i de nye omgivelser) var et godt mål for hvor længe personerne havde levet i de civiliserede omgivelser og at det derfor måtte være interessant at se hvor godt x_1 kunne forklare blodtrykket y. Første skridt er derfor at fitte en simpel lineær regressionsmodel med x_1 som forklarende variabel. Man finder den estimerede regressionslinie til

$$y = 134 - 16x_1 ,$$

og det tilhørende variansskøn er 163 med 37 frihedsgrader.

Hvis man i et koordinatsystem afsætter y mod x_1 viser det sig imidlertid, se Figur 12.3, at det bestemt ikke virker særlig rimeligt at hævde at (middelværdien af) y afhænger lineært af x_1. Derfor må man give sig til at overveje om andre af de målte baggrundsvariable med fordel kan inddrages.

Nu ved man, at en persons vægt har betydning for den pågældendes blodtryk, så næste modelforslag er en multipel regressionsmo-

del med både x_1 og x_2 som forklarende variable. Skønnene over parametrene β_0, β_1 og β_2 i regressionsligningen

$$y = \beta_0 + x_1\beta_1 + x_2\beta_2$$

bestemmes (jf. side 298) som løsning til ligningerne

$$\begin{array}{rcrcrcr} 39\,\beta_0 & + & 15.066\,\beta_1 & + & 2463.20\,\beta_2 & = & 4969 \\ 15.066\,\beta_0 & + & 7.826896\,\beta_1 & + & 969.7395\,\beta_2 & = & 1887.944 \\ 2463.20\,\beta_0 & + & 969.7395\,\beta_1 & + & 157488.16\,\beta_2 & = & 315680.8 \end{array}$$

Man finder at $\widehat{\beta}_0 = 60.8775$, $\widehat{\beta}_1 = -26.78738$ og $\widehat{\beta}_2 = 1.21726$, så den estimerede regressionsligning er

$$y = 61 - 27x_1 + 1.2x_2\,, \qquad (12.18)$$

og skønnet over variansen bliver denne gang 96 med 36 frihedsgrader.

Det ses at ved at inddrage x_2 er variansen gået drastisk ned, fra 163 til 96. Deraf kan man dog ikke slutte at (12.18) er en god beskrivelse af data, kun at den er bedre end den forrige. Man bør undersøge residualerne for at kunne vurdere modellens kvalitet – det vil vi dog ikke gøre her.

Hvis man har ladet en computer foretage udregningerne, vil man sandsynligvis også have fået oplyst parameterskønnenes middelfejl og fået at vide om parametrene hver især var signifikant forskellige fra 0. I det konkrete tilfælde vil man da have fået at vide, at når man kun bruger x_1 så er koefficienten til x_1 ikke signifikant forskellig fra 0, men når man benytter både x_1 og x_2 så er alle koefficienter signifikant forskellige fra 0. Det kan man fortolke på den måde, at blodtrykket afhænger signifikant af både x_1 og x_2, således at jo længere man har levet i de nye omgivelser jo lavere blodtryk, og jo større vægt jo højere blodtryk; men da det nok også er sådan at jo længere tid man har boet i "civilisationen" jo mere vejer man, så vil de to virkninger udjævne hinanden hvis man ikke sørger for at inddrage begge forklarende variable. □

Liste over eksempler

Liste over resumeer

Liste over anvendte symboler

Symboler der begynder med bogstaver

A	betegner tit en hændelse, dvs. en delmængde af udfaldsrummet.
B	$B(x, y)$: betafunktionen.
cos $\cos^{-1}$	cos er *cosinus*-funktionen, $\cos^{-1}$ er dens inverse funktion (der også kaldes arcus-cosinus), dvs. $x = \cos^{-1}(y)$ hvis og kun hvis $y = \cos x$ og $0 \leq x < 2\pi$.
Cov	*Kovarians*. $\mathrm{Cov}(X, Y)$ betegner kovariansen mellem de to stokastiske variable X og Y.
$\frac{d}{dx}$	*Differentiation*: Hvis $y = g(x)$ er en differentiabel funktion af x, så er $\frac{dy}{dx}$ og $\frac{d}{dx}g$ to betegnelser for differentialkvotienten af g; en tredje betegnelse er g'.
$\frac{\partial}{\partial x_j}$	*Partiel differentiation*: Hvis g er en funktion af de to variable x_1 og x_2, så betegner $\frac{\partial}{\partial x_1}g$ den afledede af funktionen $x_1 \mapsto g(x_1, x_2)$.
e	grundtallet for den naturlige logaritme; $e \approx 2.7183$. e kan også betegne et *residual*.
E	*Middelværdi*: $\mathrm{E}\,X$ betegner middelværdien af den stokastiske variabel X.

exp	*eksponentialfunktionen*: $\exp(x) = e^x$. Der gælder at $y = \exp(x)$ hvis og kun hvis $x = \ln y$.
f	Hvis f betegner et tal er det ofte et antal *frihedsgrader*, hvis f betegner en funktion er det ofte en *sandsynlighedsfunktion* eller en *sandsynlighedstæthedsfunktion*.
F	betegner ofte en *F-teststørrelse*. F kan også betegne en (kumuleret) (sandsynligheds-)fordelingsfunktion.
F_{f_1,f_2}	betegner en stokastisk variabel som følger F-fordelingen med f_1 og f_2 frihedsgrader.
H	statistiske *hypoteser* plejer at blive benævnt H.
L	betegner som regel en *likelihoodfunktion*.
ln	den *naturlige logaritme*, $\ln x = \int_1^x \frac{1}{t}\,dt$. Funktionen ln har den egenskab at $$\ln(a\,b) \;=\; \ln a + \ln b\,.$$
$\ln L$	*log-likelihoodfunktionen*, dvs. logaritmen til likelihoodfunktionen L.
$\mathbf{N}$	betegner mængden af *naturlige tal*, dvs. $\mathbf{N} = \{1, 2, 3, \ldots\}$.
$\mathbf{N}_0$	betegner mængden af *hele ikke-negative tal*, $\mathbf{N}_0 = \{0, 1, 2, 3, \ldots\}$.
$\mathcal{N}(\mu\,,\,\sigma^2)$	*normalfordelingen* med parametre μ og σ^2.
obs	Q_{obs}, t_{obs} osv. betegner den *observerede værdi* af den stokastiske variabel Q, t osv.
$\mathrm{P}(A)$	*sandsynligheden for* hændelsen A.
$\mathrm{P}(X \in A)$	sandsynligheden for at den stokastiske variabel X tilhører A.
$\mathrm{P}(\ldots \mid \ldots)$	*betinget sandsynlighed*: $\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}$ er den betingede sandsynlighed for A givet at B indtræffer.

$\mathbf{R}$	betegner mængden af *reelle tal*, $\mathbf{R} =]-\infty, +\infty[$.
$\mathbf{R}^n$	betegner mængden af talsæt $(x_1, x_2, \ldots, x_n)$ hvor hvert x_j tilhører mængden $\mathbf{R}$ af reelle tal.
SP	*Sum af Produkter* af afvigelser fra gennemsnittet: $$SP_{xy} = \sum(x - \overline{x})(y - \overline{y}) .$$
SS	*Sum af Kvadrater* af afvigelser fra gennemsnittet: $$SS_x = \sum(x - \overline{x})^2 .$$
s^2	betegner ofte skønnet over variansen σ^2 i normalfordelingen.
sin $\sin^{-1}$	sin er *sinus*-funktionen, $\sin^{-1}$ er dens inverse funktion (der også kaldes arcus-sinus), dvs. $x = \sin^{-1}(y)$ hvis og kun hvis $y = \sin x$ og $0 \leq x < 2\pi$.
t	t kan betegne en *t-teststørrelse*, t kan også betegne en *tid*, og endelig kan t betegne en *transformation*, dvs. en afbildning. Hvis t er en transformation der afbilder punkter i $\mathcal{X}$ over i $\mathcal{Y}$, skrives $t : \mathcal{X} \to \mathcal{Y}$.
t^{-1}	kan betegne den inverse afbildning til afbildningen t, dvs. $t(x) = y$ er ensbetydende med at $x = t^{-1}(y)$.
t_f	betegner en stokastisk variabel som følger t-fordelingen med f frihedsgrader.
Var	*Varians*. Var X betegner variansen på den stokastiske variabel X.
x, x_1, x_2 osv.	udfald (dvs. elementer i $\mathcal{X}$).
x_α	betegner undertiden en *α-fraktil*.
$\boldsymbol{x}$	$= (x_1, x_2, \ldots, x_n)$.
X	stokastisk variabel; antager værdier $x, x_1, x_2 \ldots$
$\boldsymbol{X}$	$= (X_1, X_2, \ldots, X_n)$.

$\mathcal{X}$	udfaldsrum (dvs. en mængde af udfald).
$\mathcal{X}_y$	niveaumængde af typen $\mathcal{X}_y = \{x \in \mathcal{X} : t(x) = y\}$.
Y	stokastisk variabel; antager værdier $y, y_1, y_2 \ldots$
y, y_1, y_2 osv.	udfald (dvs. elementer i $\mathcal{Y}$).
$\mathcal{Y}$	udfaldsrum (dvs. mængder af udfald).

Symboler med græske bogstaver

α	græsk *alfa*
β	græsk *beta*
γ	græsk *gamma*
Γ	græsk *stort gamma*, jf. gamma-funktionen.
ε	græsk *epsilon*, betegner ofte *testsandsynligheden*
λ	græsk *lambda*. Betegner undertiden parameteren i en Poisson-fordeling eller intensiteten i en Poissonproces.
μ	græsk *my*, betegner ofte en middelværdi.
π	græsk *pi*, forholdet mellem en cirkels omkreds og diameter. $\pi \approx 3.14159$.
σ	græsk *sigma*. Betegner ofte standardafvigelse.
σ^2	Ofte variansparameteren i en normalfordeling.
ϕ	græsk *fi*, betegner ofte tæthedsfunktionen i den normerede normalfordeling $\mathcal{N}(0\,,\,1)$: $$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}x^2) \,.$$
χ	græsk *khi*
χ^2	*khi i anden*, jf. χ^2-fordeling.

χ^2_f betegner en stokastisk variabel som er χ^2-fordelt med f frihedsgrader.

$\sum$ græsk *stort sigma*, summationstegn:

- $\sum_{i=1}^{m} f(x_i)$ eller $\sum_{i=1}^{m} f(x_i)$ betyder $f(x_1) + f(x_2) + \ldots + f(x_m)$.
- $\sum_{x \in A} f(x)$ eller $\sum_{x \in A} f(x)$ betyder summen af de funktionsværdier $f(x)$ for hvilke x tilhører mængden A.
- $\sum_{x \leq a} f(x)$ betyder summen af de funktionsværdier $f(x)$ for hvilke x er mindre end eller lig a.
- Man nøjes undertiden med at skrive $\sum f(x)$, idet det så er underforstået hvilken mængde af x-er der summeres over.

$\prod$ græsk *stort pi*, produkttegn:

- $\prod_{i=1}^{m} f(x_i)$ eller $\prod_{i=1}^{m} f(x_i)$ betyder $f(x_1) \times f(x_2) \times \ldots \times f(x_m)$.
- $\prod_{x \in A} f(x)$ eller $\prod_{x \in A} f(x)$ betyder produktet af de funktionsværdier $f(x)$ for hvilke x tilhører mængden A.
- $\prod_{x \leq a} f(x)$ betyder produktet af de funktionsværdier $f(x)$ for hvilke x er mindre end eller lig a.

Symboler der ikke er bogstaver

$\rightarrow$

1. $x \rightarrow +\infty$ betyder at x går mod $+\infty$.
2. $t : \mathcal{X} \rightarrow \mathcal{Y}$ betyder at t afbilder punkterne i mængden $\mathcal{X}$ over i mængden $\mathcal{Y}$.

$\mapsto$ $t : x \mapsto x^2$ betyder at t er afbildningen der afbilder et tal x over i tallet x^2.

$\searrow$ $h \searrow 0$ betyder at h går mod 0 gennem positive værdier.

$'$ *Differentiation.* Hvis g er en funktion, så er g' differentialkvotienten af g (også kaldet den afledede af g).

$\sim$ *er fordelt som.*

Eks.: $X \sim \mathcal{N}(0, 1)$ betyder at den stokastiske variabel X følger $\mathcal{N}(0, 1)$-fordelingen.

$\approx$ *er omtrent lig med*

$\times$ *gangetegn.* Tegnet benyttes også ved dannelse af *produktmængde*, f.eks. er $\mathcal{X}_1 \times \mathcal{X}_2$ mængden af par (x_1, x_2) hvor $x_1 \in \mathcal{X}_1$ og $x_2 \in \mathcal{X}_2$.

$!$ *fakultetsfunktion.* $n! = 1 \times 2 \times 3 \times \ldots \times n$.

$| \; |$ $|a|$ betegner den *numeriske værdi* eller *absolutte værdi* af tallet a, dvs.

$$|a| = \begin{cases} a & \text{hvis } a \geq 0 \\ -a & \text{ellers.} \end{cases}$$

$\cdot$ Et *punkt* på et index's plads betyder at der er summeret over det pågældende index.

Eks.: Hvis der er tale om tal $x_1, x_2, \ldots, x_n$, så betyder $x_\cdot$ summen af disse x-er, dvs.

$$x_\cdot = \sum_{i=1}^{n} x_i \; .$$

Hvis der er tale om tal x_{ij} hvor i går fra 1 til r og j fra 1 til s, så er

$$x_{i\cdot} = \sum_{j=1}^{s} x_{ij} \; ,$$

$$x_{\cdot j} = \sum_{i=1}^{r} x_{ij} \; ,$$

$$x_{\cdot\cdot} = \sum_{i=1}^{r} \sum_{j=1}^{s} x_{ij} \; .$$

$\bar{}$ *streg*. Betegner gennemsnit: $\overline{x}$ er gennemsnittet af x-erne. *streg*-notationen kombineres ofte med *punkt*-notationen for at betegne gennemsnit over et bestemt index.

Eks.: Hvis der er tale om tal x_{ij} hvor i går fra 1 til r og j fra 1 til s, så betegnes gennemsnittet over j, dvs. gennemsnittet i den i-te række,

$$\begin{aligned} \overline{x}_{i\bullet} &= x_{i\bullet}/s \\ &= \frac{1}{s}\sum_{j=1}^{s} x_{ij} \ , \end{aligned}$$

gennemsnittet over i, dvs. gennemsnittet i den j-te søjle, betegnes

$$\begin{aligned} \overline{x}_{\bullet j} &= x_{\bullet j}/r \\ &= \frac{1}{r}\sum_{i=1}^{r} x_{ij} \ , \end{aligned}$$

og totalgennemsnittet betegnes

$$\begin{aligned} \overline{x} &= \overline{x}_{\bullet\bullet} \\ &= x_{\bullet\bullet}/rs \\ &= \frac{1}{rs}\sum_{i=1}^{r}\sum_{j=1}^{s} x_{ij} \ . \end{aligned}$$

$\hat{}$ *hat*. En *hat* over et parameternavn betegner *maksimaliseringsestimatet* for den pågældende parameter.

Hvis man også estimerer parameteren under forskellige hypoteser, kan de tilsvarende estimater betegnes med flere hatter, f.eks. $\widehat{\widehat{\alpha}}$. Man kan også bruge en tilde: $\widetilde{\alpha}$.

$\tilde{}$ *tilde*.

(Se ovenfor.)

$\binom{n}{x}$ *binomialkoefficient*.

$\binom{n}{x_1\ x_2\ \ldots\ x_k}$ *multinomialkoefficient*

$\{\ \}$ *mængde-parenteser*:

$\{x_1, x_2, \ldots, x_m\}$ betegner mængden bestående af elementerne $x_1, x_2, \ldots, x_m$.

$\{x : f(x) \leq 3\}$ betegner mængden af x-er (fra det aktuelle udfaldsrum) for hvilke betingelsen $f(x) \leq 3$ er opfyldt.

∞ *uendelig*

$+\infty\ -\infty$ *plus uendelig* og *minus uendelig*; den reelle akse går fra $-\infty$ til $+\infty$.

Stikordsregister

1/78 "TANKER OM EN PRAKSIS" - et matematikprojekt.
Projektrapport af: Anne Jensen, Lena Lindenskov, Marianne Kesselhahn og Nicolai Lomholt.
Vejleder: Anders Madsen

2/78 "OPTIMERING" - Menneskets forøgede beherskelsesmuligheder af natur og samfund.
Projektrapport af: Tom J. Andersen, Tommy R. Andersen, Gert Krenøe og Peter H. Lassen
Vejleder: Bernhelm Boss.

3/78 "OPCAVESAMLING", breddekursus i fysik.
Af: Lasse Rasmussen, Aage Bonde Kræmmer og Jens Højgaard Jensen.

4/78 "TRE ESSAYS" - om matematikundervisning, matematiklæreruddannelsen og videnskabsrindalismen.
Af: Mogens Niss
Nr. 4 er p.t. udgået.

5/78 "BIBLIOGRAFISK VEJLEDNING til studiet af DEN MODERNE FYSIKS HISTORIE".
Af: Helge Kragh.
Nr. 5 er p.t. udgået.

6/78 "NOGLE ARTIKLER OG DEBATINDLÆG OM - læreruddannelse og undervisning i fysik, og - de naturvidenskabelige fags situation efter studenteroprøret".
Af: Karin Beyer, Jens Højgaard Jensen og Bent C. Jørgensen.

7/78 "MATEMATIKKENS FORHOLD TIL SAMFUNDSØKONOMIEN".
Af: B.V. Gnedenko.
Nr. 7 er udgået.

8/78 "DYNAMIK OG DIAGRAMMER". Introduktion til energy-bond-graph formalismen.
Af: Peder Voetmann Christiansen.

9/78 "OM PRAKSIS' INDFLYDELSE PÅ MATEMATIKKENS UD - VIKLING". - Motiver til Kepler's: "Nova Stereometria Doliorum Vinariom".
Projektrapport af: Lasse Rasmussen .
Vejleder: Anders Madsen.

10/79 "TERMODYNAMIK I GYMNASIET".
Projektrapport af: Jan Christensen og Jeanne Mortensen,
Vejledere: Karin Beyer og Peder Voetmann Christiansen.

11/79 "STATISTISKE MATERIALER".
Af: Jørgen Larsen.

12/79 "LINEÆRE DIFFERENTIALLIGNINGER OG DIFFERENTIALLIGNINGSSYSTEMER".
Af: Mogens Brun Heefelt.
Nr. 12 er udgået.

13/79 "CAVENDISH'S FORSØG I GYMNASIET".
Projektrapport af: Gert Kreinøe.
Vejleder: Albert Chr. Paulsen.

14/79 "BOOKS ABOUT MATHEMATICS: History, Philosophy, Education, Models, System Theory, and Works of".
Af: Else Høyrup.
Nr. 14 er p.t. udgået.

15/79 "STRUKTUREL STABILITET OG KATASTROFER i systemer i og udenfor termodynamisk ligevægt".
Specialeopgave af: Leif S. Striegler.
Vejleder: Peder Voetmann Ckristiansen.

16/79 "STATISTIK I KRÆFTFORSKNINGEN".
Projektrapport af: Michael Olsen og Jørn Jensen.
Vejleder: Jørgen Larsen.

17/79 "AT SPØRGE OG AT SVARE i fysikundervisningen".
Af: Albert Christian Paulsen.

18/79 "MATHEMATICS AND THE REAL WORLD", Proceedings af an International Workshop, Roskilde University Centre, Denmark, 1978.
Preprint.
Af: Bernhelm Booss og Mogens Niss (eds.)

19/79 "GEOMETRI, SKOLE OG VIRKELIGHED".
Projektrapport af: Tom J. Andersen, Tommy R. Andersen og Per H.H. Larsen.
Vejleder: Mogens Niss.

20/79 "STATISTISKE MODELLER TIL BESTEMMELSE AF SIKRE DOSER FOR CARCINOGENE STOFFER".
Projektrapport af: Michael Olsen og Jørn Jensen.
Vejleder: Jørgen Larsen

21/79 "KONTROL I GYMNASIET-FORMÅL OG KONSEKVENSER".
Projektrapport af: Crilles Bacher, Per S.Jensen, Preben Jensen og Torben Nysteen.

22/79 "SEMIOTIK OG SYSTEMEGENSKABER (1)".
1-port lineært response og støj i fysikken.
Af: Peder Voetmann Christiansen.

23/79 "ON THE HISTORY AF EARLY WAVE MECHANICS - with special emphasis on the role af realitivity".
Af: Helge Kragh.

24/80 a+b "MATEMATIKOPFATTELSER HOS 2.G'ERE".
1. En analyse. 2. Interviewmateriale.
Projektrapport af: Jan Christensen og Knud Lindhardt Rasmussen.
Vejleder: Mogens Niss.

25/80 "EKSAMENSOPGAVER", Dybdemodulet/fysik 1974-79.

26/80 "OM MATEMATISKE MODELLER".
En projektrapport og to artikler.
Af: Jens Højgaard Jensen m.fl.

27/80 "METHODOLOGY AND PHILOSOPHY AF SCIENCE IN PAUL DIRAC's PHYSICS".
Af: Helge Kragh.

28/80 "DIELEKTRISK RELAXATION - et forslag til en ny model bygget på væskernes viscoelastiske egenskaber".
Projektrapport af: Gert Kreinøe.
Vejleder: Niels Boye Olsen.

29/80 "ODIN - undervisningsmateriale til et kursus i differentialligningsmodeller".
Projektrapport af: Tommy R. Andersen, Per H.H. Larsen og Peter H. Lassen.
Vejleder: Mogens Brun Heefelt.

30/80 "FUSIONSENERGIEN - - - ATOMSAMFUNDETS ENDESTATION".
Af: Oluf Danielsen.
Nr. 30 er udgået.

31/80 "VIDENSKABSTEORETISKE PROBLEMER VED UNDERVISNINGS - SYSTEMER BASERET PÅ MÆNGDELÆRE".
Projektrapport af: Troels Lange og Jørgen Karrebæk.
Vejleder: Stig Andur Pedersen.
Nr. 31 er p.t. udgået.

32/80 "POLYMERE STOFFERS VISCOELASTISKE EGENSKABER - BELYST VED HJÆLP AF MEKANISKE IMPEDANSMÅLINGER OG MÖSSBAUEREFFEKTMÅLINGER".
Projektrapport af: Crilles Bacher og Preben Jensen.
Vejledere: Niels Boye Olsen og Peder Voetmann Christiansen.

33/80 "KONSTITUERING AF FAG INDEN FOR TEKNISK - NATURVIDENSKABELIGE UDDANNELSER. I-II".
Af: Arne Jakobsen.

34/80 "ENVIRONMENTAL IMPACT AF WIND ENERGY UTILIZATION".
ENERGY SERIES NO. I.
Af: Bent Sørensen
Nr. 34 er udgået.

35/80 "HISTORISKE STUDIER I DEN NYERE ATOMFYSIKS UDVIKLING".
Af: Helge Kragh.

36/80 "HVAD ER MENINGEN MED MATEMATIKUNDERVISNINGEN?".
Fire artikler.
Af: Mogens Niss.

37/80 "RENEWABLE ENERGY AND ENERGY STORAGE".
ENERGY SERIES NO. 2.
Af: Bent Sørensen.

38/81 "TIL EN HISTORIETEORI OM NATURERKENDELSE, TEKNOLOGI OG SAMFUND".
Projektrapport af: Erik Gade, Hans Hedal, Henrik Lau og Finn Physant.
Vejledere: Stig Andur Pedersen, Helge Kragh og Ib Thiersen.
Nr. 38 er p.t. udgået.

39/81 "TIL KRITIKKEN AF VÆKSTØKONOMIEN".
Af: Jens Højgaard Jensen.

40/81 "TELEKOMMUNIKATION I DANMARK - oplæg til en teknologivurdering".
Projektrapport af: Arne Jørgensen, Bruno Petersen og Jan Vedde.
Vejleder: Per Nørgaard.

41/81 "PLANNING AND POLICY CONSIDERATIONS RELATED TO THE INTRODUCTION OF RENEWABLE ENERGY SOURCES INTO ENERGY SUPPLY SYSTEMS".
ENERGY SERIES NO. 3.
Af: Bent Sørensen.

42/81 "VIDENSKAB TEORI SAMFUND - En introduktion til materialistiske videnskabsopfattelser".
Af: Helge Kragh og Stig Andur Pedersen.

43/81 1."COMPARATIVE RISK ASSESSMENT OF TOTAL ENERGY SYSTEMS".
2."ADVANTAGES AND DISADVANTAGES OF DECENTRALIZATION".
ENERGY SERIES NO. 4.
Af: Bent Sørensen.

44/81 "HISTORISKE UNDERSØGELSER AF DE EKSPERIMENTELLE FORUDSÆTNINGER FOR RUTHERFORDS ATOMMODEL".
Projektrapport af: Niels Thor Nielsen.
Vejleder: Bent C. Jørgensen.

45/82 Er aldrig udkommet.

46/82 "EKSEMPLARISK UNDERVISNING OG FYSISK ERKENDESE-
1+11 ILLUSTRERET VED TO EKSEMPLER".
Projektrapport af: Torben O.Olsen, Lasse Rasmussen og Niels Dreyer Sørensen.
Vejleder: Bent C. Jørgensen.

47/82 "BARSEBÄCK OG DET VÆRST OFFICIELT-TÆNKELIGE UHELD".
ENERGY SERIES NO. 5.
Af: Bent Sørensen.

48/82 "EN UNDERSØGELSE AF MATEMATIKUNDERVISNINGEN PÅ ADGANGSKURSUS TIL KØBENHAVNS TEKNIKUM".
Projektrapport af: Lis Eilertzen, Jørgen Karrebæk, Troels Lange, Preben Nørregaard, Lissi Pedesen, Laust Rishøj, Lill Røn og Isac Showiki.
Vejleder: Mogens Niss.

49/82 "ANALYSE AF MULTISPEKTRALE SATELLITBILLEDER".
Projektrapport af: Preben Nørregaard.
Vejledere: Jørgen Larsen og Rasmus Ole Rasmussen.

50/82 "HERSLEV - MULIGHEDER FOR VEDVARENDE ENERGI I EN LANDSBY".
ENERGY SERIES NO. 6.
Rapport af: Bent Christensen, Bent Hove Jensen, Dennis B. Møller, Bjarne Laursen, Bjarne Lillethorup og Jacob Mørch Pedersen.
Vejleder: Bent Sørensen.

51/82 "HVAD KAN DER GØRES FOR AT AFHJÆLPE PIGERS BLOKERING OVERFOR MATEMATIK ?"
Projektrapport af: Lis Eilertzen, Lissi Pedersen, Lill Røn og Susanne Stender.

52/82 "DESUSPENSION OF SPLITTING ELLIPTIC SYMBOLS".
Af: Bernhelm Booss og Krzysztof Wojciechowski.

53/82 "THE CONSTITUTION OF SUBJECTS IN ENGINEERING EDUCATION".
Af: Arne Jacobsen og Stig Andur Pedersen.

54/82 "FUTURES RESEARCH" - A Philosophical Analysis of Its Subject-Matter and Methods.
Af: Stig Andur Pedersen og Johannes Witt-Hansen.

55/82 "MATEMATISKE MODELLER" - Litteratur på Roskilde Universitetsbibliotek.
En biografi.
Af: Else Høyrup.

Vedr. tekst nr. 55/82 se også tekst nr. 62/83.

56/82 "ÉN - TO - MANGE" -
En undersøgelse af matematisk økologi.
Projektrapport af: Troels Lange.
Vejleder: Anders Madsen.

57/83 "ASPECT EKSPERIMENTET"-
Skjulte variable i kvantemekanikken?
Projektrapport af: Tom Juul Andersen.
Vejleder: Peder Voetmann Christiansen.
Nr. 57 er udgået.

58/83 "MATEMATISKE VANDRINGER" - Modelbetragtninger over spredning af dyr mellem småbiotoper i agerlandet.
Projektrapport af: Per Hammershøj Jensen og Lene Vagn Rasmussen.
Vejleder: Jørgen Larsen.

59/83 "THE METHODOLOGY OF ENERGY PLANNING".
ENERGY SERIES NO. 7.
Af: Bent Sørensen.

60/83 "MATEMATISK MODEKSPERTISE"- et eksempel.
Projektrapport af: Erik O. Gade, Jørgen Karrebæk og Preben Nørregaard.
Vejleder: Anders Madsen.

61/83 "FYSIKS IDEOLOGISKE FUNKTION, SOM ET EKSEMPEL PÅ EN NATURVIDENSKAB - HISTORISK SET".
Projektrapport af: Annette Post Nielsen.
Vejledere: Jens Høyrup, Jens Højgaard Jensen og Jørgen Vogelius.

62/83 "MATEMATISKE MODELLER" - Litteratur på Roskilde Universitetsbibliotek.
En biografi 2. rev. udgave.
Af: Else Høyrup.

63/83 "GREATING ENERGY FUTURES:A SHORT GUIDE TO ENERGY PLANNING".
ENERGY SERIES No. 8.
Af: David Crossley og Bent Sørensen.

64/83 "VON MATEMATIK UND KRIEG".
Af: Berhelm Booss og Jens Høyrup.

65/83 "ANVENDT MATEMATIK - TEORI ELLER PRAKSIS".
Projektrapport af: Per Hedegård Andersen, Kirsten Habekost, Carsten Holst-Jensen, Annelise von Moos, Else Marie Pedersen og Erling Møller Pedersen.
Vejledere: Bernhelm Booss og Klaus Grünbaum.

66/83 "MATEMATISKE MODELLER FOR PERIODISK SELEKTION I ESCHERICHIA COLI".
Projektrapport af: Hanne Lisbet Andersen, Ole Richard Jensen og Klavs Frisdahl.
Vejledere: Jørgen Larsen og Anders Hede Madsen.

67/83 "ELEPSOIDE METODEN - EN NY METODE TIL LINEÆR PROGRAMMERING?"
Projektrapport af: Lone Biilmann og Lars Boye.
Vejleder: Mogens Brun Heefelt.

68/83 "STOKASTISKE MODELLER I POPULATIONSGENETIK"
- til kritikken af teoriladede modeller.
Projektrapport af: Lise Odgård Gade, Susanne Hansen, Michael Hviid og Frank Mølgård Olsen.
Vejleder: Jørgen Larsen.

69/83 "ELEVFORUDSÆTNINGER I FYSIK"
- en test i 1.g med kommentarer.
Af: Albert C. Paulsen.

70/83 "INDLÆRINGS - OG FORMIDLINGSPROBLEMER I MATEMATIK PÅ VOKSENUNDERVISNINGSNIVEAU".
Projektrapport af: Hanne Lisbet Andersen, Torben J. Andreasen, Svend Åge Houmann, Helle Glerup Jensen, Keld Fl. Nielsen, Lene Vagn Rasmussen.
Vejleder: Klaus Grünbaum og Anders Hede Madsen.

71/83 "PIGER OG FYSIK"
- et problem og en udfordring for skolen?
Af: Karin Beyer, Sussanne Blegaa, Birthe Olsen, Jette Reich og Mette Vedelsby.

72/83 "VERDEN IFØLGE PEIRCE" - to metafysiske essays, om og af C.S Peirce.
Af: Peder Voetmann Christiansen.

73/83 ""EN ENERGIANALYSE AF LANDBRUG"
- økologisk contra traditionelt.
ENERGY SERIES NO. 9
Specialeopgave i fysik af: Bent Hove Jensen,
Vejleder: Bent Sørensen.

74/84 "MINIATURISERING AF MIKROELEKTRONIK" - om videnskabeliggjort teknologi og nytten af at lære fysik.
Projektrapport af: Bodil Harder og Linda Szkotak Jensen.
Vejledere: Jens Højgaard Jensen og Bent C. Jørgensen.

75/84 "MATEMATIKUNDERVISNINGEN I FREMTIDENS GYMNASIUM"
- Case: Lineær programmering.
Projektrapport af: Morten Blomhøj, Klavs Frisdahl og Frank Mølgaard Olsen.
Vejledere: Mogens Brun Heefelt og Jens Bjørneboe.

76/84 "KERNEKRAFT I DANMARK?" - Et høringssvar indkaldt af miljøministeriet, med kritik af miljøstyrelsens rapporter af 15. marts 1984.
ENERGY SERIES No. 10
Af: Niels Boye Olsen og Bent Sørensen.

77/84 "POLITISKE INDEKS - FUP ELLER FAKTA?"
Opinionsundersøgelser belyst ved statistiske modeller.
Projektrapport af: Svend Åge Houmann, Keld Nielsen og Susanne Stender.
Vejledere: Jørgen Larsen og Jens Bjørneboe.

78/84 "JÆVNSTRØMSLEDNINGSEVNE OG GITTERSTRUKTUT I AMORFT GERMANIUM".
Specialrapport af: Hans Hedal, Frank C. Ludvigsen og Finn C. Physant.
Vejleder: Niels Boye Olsen.

79/84 "MATEMATIK OG ALMENDANNELSE".
Projektrapport af: Henrik Coster, Mikael Wennerberg Johansen, Povl Kattler, Birgitte Lydholm og Morten Overgaard Nielsen.
Vejleder: Bernhelm Booss.

80/84 "KURSUSMATERIALE TIL MATEMATIK B".
Af: Mogens Brun Heefelt.

81/84 "FREKVENSAFHÆNGIG LEDNINGSEVNE I AMORFT GERMANIUM".
Specialerapport af: Jørgen Wind Petersen og Jan Christensen.
Vejleder: Niels Boye Olsen.

82/84 "MATEMATIK - OG FYSIKUNDERVISNINGEN I DET AUTOMATISEREDE SAMFUND".
Rapport fra et seminar afholdt i Hvidovre 25-27 april 1983.
Red.: Jens Højgaard Jensen, Bent C. Jørgensen og Mogens Niss.

83/84 "ON THE QUANTIFICATION OF SECURITY":
PEACE RESEARCH SERIES NO. 1
Af: Bent Sørensen
nr. 83 er p.t. udgået

84/84 "NOGLE ARTIKLER OM MATEMATIK, FYSIK OG ALMENDANNELSE".
Af: Jens Højgaard Jensen, Mogens Niss m. fl.

85/84 "CENTRIFUGALREGULATORER OG MATEMATIK".
Specialerapport af: Per Hedegård Andersen, Carsten Holst-Jensen, Else Marie Pedersen og Erling Møller Pedersen.
Vejleder: Stig Andur Pedersen.

86/84 "SECURITY IMPLICATIONS OF ALTERNATIVE DEFENSE OPTIONS FOR WESTERN EUROPE".
PEACE RESEARCH SERIES NO. 2
Af: Bent Sørensen.

87/84 "A SIMPLE MODEL OF AC HOPPING CONDUCTIVITY IN DISORDERED SOLIDS".
Af: Jeppe C. Dyre.

88/84 "RISE, FALL AND RESURRECTION OF INFINITESIMALS".
Af: Detlef Laugwitz.

89/84 "FJERNVARMEOPTIMERING".
Af: Bjarne Lillethorup og Jacob Mørch Pedersen.

90/84 "ENERGI I 1.G - EN TEORI FOR TILRETTELÆGGELSE".
Af: Albert Chr. Paulsen.

91/85 "KVANTETEORI FOR GYMNASIET".
1. Lærervejledning
Projektrapport af: Biger Lundgren, Henning Sten Hansen og John Johansson.
Vejleder: Torsten Meyer.

92/85 "KVANTETEORI FOR GYMNASIET".
2. Materiale
Projektrapport af: Biger Lundgren, Henning Sten Hansen og John Johansson.
Vejleder: Torsten Meyer.

93/85 "THE SEMIOTICS OF QUANTUM - NON - LOCALITY".
Af: Peder Voetmann Christiansen.

94/85 "TREENIGHEDEN BOURBAKI - generalen, matematikeren og ånden".
Projektrapport af: Morten Blomhøj, Klavs Frisdahl og Frank M. Olsen.
Vejleder: Mogens Niss.

95/85 "AN ALTERNATIV DEFENSE PLAN FOR WESTERN EUROPE".
PEACE RESEARCH SERIES NO. 3
Af: Bent Sørensen

96/85 "ASPEKTER VED KRAFTVARMEFORSYNING".
Af: Bjarne Lilletorup.
Vejleder: Bent Sørensen.

97/85 "ON THE PHYSICS OF A.C. HOPPING CONDUCTIVITY".
Af: Jeppe C. Dyre.

98/85 "VALGMULIGHEDER I INFORMATIONSALDEREN".
Af: Bent Sørensen.

99/85 "Der er langt fra Q til R".
Projektrapport af: Niels Jørgensen og Mikael Klintorp.
Vejleder: Stig Andur Pedersen.

100/85 "TALSYSTEMETS OPBYGNING".
Af: Mogens Niss.

101/85 "EXTENDED MOMENTUM THEORY FOR WINDMILLS IN PERTURBATIVE FORM".
Af: Ganesh Sengupta.

102/85 OPSTILLING OG ANALYSE AF MATEMATISKE MODELLER, BELYST VED MODELLER OVER KØERS FODEROPTAGELSE OG - OMSÆTNING".
Projektrapport af: Lis Eilertzen, Kirsten Habekost, Lill Røn og Susanne Stender.
Vejleder: Klaus Grünbaum.

103/85 "ØDSLE KOLDKRIGERE OG VIDENSKABENS LYSE IDEER".
Projektrapport af: Niels Ole Dam og Kurt Jensen.
Vejleder: Bent Sørensen.

104/85 "ANALOGREGNEMASKINEN OG LORENZLIGNINGER".
Af: Jens Jæger.

105/85 "THE FREQUENCY DEPENDENCE OF THE SPECIFIC HEAT AF THE GLASS REANSITION".
Af: Tage Christensen.

"A SIMPLE MODEL AF AC HOPPING CONDUCTIVITY".
Af: Jeppe C. Dyre.
Contributions to the Third International Conference on the Structure of Non - Crystalline Materials held in Grenoble July 1985.

106/85 "QUANTUM THEORY OF EXTENDED PARTICLES".
Af: Bent Sørensen.

107/85 "EN MYG GØR INGEN EPIDEMI".
- flodblindhed som eksempel på matematisk modellering af et epidemiologisk problem.
Projektrapport af: Per Hedegård Andersen, Lars Boye, CarstenHolst Jensen, Else Marie Pedersen og Erling Møller Pedersen.
Vejleder: Jesper Larsen.

108/85 "APPLICATIONS AND MODELLING IN THE MATEMATICS CURRICULUM" - state and trends -
Af: Mogens Niss.

109/85 "COX I STUDIETIDEN" - Cox's regressionsmodel anvendt på studenteroplysninger fra RUC.
Projektrapport af: Mikael Wennerberg Johansen, Poul Kattler og Torben J. Andreasen.
Vejleder: Jørgen Larsen.

110/85 "PLANNING FOR SECURITY".
Af: Bent Sørensen

111/85 JORDEN RUNDT PÅ FLADE KORT".
Projektrapport af: Birgit Andresen, Beatriz Quinones og Jimmy Staal.
Vejleder: Mogens Niss.

112/85 "VIDENSKABELIGGØRELSE AF DANSK TEKNOLOGISK INNOVATION FREM TIL 1950 - BELYST VED EKSEMPLER".
Projektrapport af: Erik Odgaard Gade, Hans Hedal, Frank C. Ludvigsen, Annette Post Nielsen og Finn Physant.
Vejleder: Claus Bryld og Bent C. Jørgensen.

113/85 "DESUSPENSION OF SPLITTING ELLIPTIC SYMBOLS II".
Af: Bernhelm Booss og Krzysztof Wojciechowski.

114/85 "ANVENDELSE AF GRAFISKE METODER TIL ANALYSE AF KONTIGENSTABELLER".
Projektrapport af: Lone Biilmann, Ole R. Jensen og Anne-Lise von Moos.
Vejleder: Jørgen Larsen.

115/85 "MATEMATIKKENS UDVIKLING OP TIL RENÆSSANCEN".
Af: Mogens Niss.

116/85 "A PHENOMENOLOGICAL MODEL FOR THE MEYER-NELDEL RULE".
Af: Jeppe C. Dyre.

117/85 "KRAFT & FJERNVARMEOPTIMERING"
Af: Jacob Mørch Pedersen.
Vejleder: Bent Sørensen

118/85 "TILFÆLDIGHEDEN OG NØDVENDIGHEDEN IFØLGE PEIRCE OG FYSIKKEN".
Af: Peder Voetmann Christiansen

119/86 "DET ER GANSKE VIST - - EUKLIDS FEMTE POSTULAT KUNNE NOK SKABE RØRE I ANDEDAMMEN".
Af: Iben Maj Christiansen
Vejleder: Mogens Niss.

120/86 "ET ANTAL STATISTISKE STANDARDMODELLER".
Af: Jørgen Larsen

121/86 "SIMULATION I KONTINUERT TID".
Af: Peder Voetmann Christiansen.

122/86 "ON THE MECHANISM OF GLASS IONIC CONDUCTIVITY".
Af: Jeppe C. Dyre.

123/86 "GYMNASIEFYSIKKEN OG DEN STORE VERDEN".
Fysiklærerforeningen, IMFUFA, RUC.

124/86 "OPGAVESAMLING I MATEMATIK".
Samtlige opgaver stillet i tiden 1974-jan. 1986.

125/86 "UVBY,β - systemet - en effektiv fotometrisk spektralklassifikation af B-,A- og F-stjerner".
Projektrapport af: Birger Lundgren.

126/86 "OM UDVIKLINGEN AF DEN SPECIELLE RELATIVITETSTEORI".
Projektrapport af: Lise Odgaard & Linda Szkotak Jensen
Vejledere: Karin Beyer & Stig Andur Pedersen.

127/86 "GALOIS' BIDRAG TIL UDVIKLINGEN AF DEN ABSTRAKTE ALGEBRA".
Projektrapport af: Pernille Sand, Heine Larsen & Lars Frandsen.
Vejleder: Mogens Niss.

128/86 "SMÅKRYB" - om ikke-standard analyse.
Projektrapport af: Niels Jørgensen & Mikael Klintorp.
Vejleder: Jeppe Dyre.

129/86 "PHYSICS IN SOCIETY"
Lecture Notes 1983 (1986)
Af: Bent Sørensen

130/86 "Studies in Wind Power"
Af: Bent Sørensen

131/86 "FYSIK OG SAMFUND" - Et integreret fysik/historieprojekt om naturanskuelsens historiske udvikling og dens samfundsmæssige betingethed.
Projektrapport af: Jakob Heckscher, Søren Brønd, Andy Wierød.
Vejledere: Jens Høyrup, Jørgen Vogelius, Jens Højgaard Jensen.

132/86 "FYSIK OG DANNELSE"
Projektrapport af: Søren Brønd, Andy Wierød.
Vejledere: Karin Beyer, Jørgen Vogelius.

133/86 "CHERNOBYL ACCIDENT: ASSESSING THE DATA.
ENERGY SERIES NO. 15.
AF: Bent Sørensen.

134/87 "THE D.C. AND THE A.C. ELECTRICAL TRANSPORT IN AsSeTe SYSTEM"
Authors: M.B.El-Den, N.B.Olsen, Ib Høst Pedersen, Petr Viščor

135/87 "INTUITIONISTISK MATEMATIKS METODER OG ERKENDELSESTEORETISKE FORUDSÆTNINGER"
MASTEMATIKSPECIALE: Claus Larsen
Vejledere: Anton Jensen og Stig Andur Pedersen

136/87 "Mystisk og naturlig filosofi: En skitse af kristendommens første og andet møde med græsk filosofi"
Projektrapport af Frank Colding Ludvigsen
Vejledere: Historie: Ib Thiersen
Fysik: Jens Højgaard Jensen

137/87 "HOPMODELLER FOR ELEKTRISK LEDNING I UORDNEDE FASTE STOFFER" - Resume af licentiatafhandling
Af: Jeppe Dyre
Vejledere: Niels Boye Olsen og Peder Voetmann Christiansen.

138/87 "JOSEPHSON EFFECT AND CIRCLE MAP."

Paper presented at The International Workshop on Teaching Nonlinear Phenomena at Universities and Schools, "Chaos in Education". Balaton, Hungary, 26 April-2 May 1987.

By: Peder Voetmann Christiansen

139/87 "Machbarkeit nichtbeherrschbarer Technik durch Fortschritte in der Erkennbarkeit der Natur"

Af: Bernhelm Booss-Bavnbek
Martin Bohle-Carbonell

140/87 "ON THE TOPOLOGY OF SPACES OF HOLOMORPHIC MAPS"

By: Jens Gravesen

141/87 "RADIOMETERS UDVIKLING AF BLODGASAPPARATUR - ET TEKNOLOGIHISTORISK PROJEKT"

Projektrapport af Finn C. Physant

Vejleder: Ib Thiersen

142/87 "The Calderón Projektor for Operators With Splitting Elliptic Symbols"

by: Bernhelm Booss-Bavnbek og
Krzysztof P. Wojciechowski

143/87 "Kursusmateriale til Matematik på NAT-BAS"

af: Mogens Brun Heefelt

144/87 "Context and Non-Locality - A Peircan Approach

Paper presented at the Symposium on the Foundations of Modern Physics The Copenhagen Interpretation 60 Years after the Como Lecture. Joensuu, Finland, 6 - 8 august 1987.

By: Peder Voetmann Christiansen

145/87 "AIMS AND SCOPE OF APPLICATIONS AND MODELLING IN MATHEMATICS CURRICULA"

Manuscript of a plenary lecture delivered at ICMTA 3, Kassel, FRG 8.-11.9.1987

By: Mogens Niss

146/87 "BESTEMMELSE AF BULKRESISTIVITETEN I SILICIUM"

- en ny frekvensbaseret målemetode.

Fysikspeciale af Jan Vedde

Vejledere: Niels Boye Olsen & Petr Viščor

147/87 "Rapport om BIS på NAT-BAS"

redigeret af: Mogens Brun Heefelt

148/87 "Naturvidenskabsundervisning med Samfundsperspektiv"

af: Peter Colding-Jørgensen DLH
Albert Chr. Paulsen

149/87 "In-Situ Measurements of the density of amorphous germanium prepared in ultra high vacuum"

by: Petr Viščor

150/87 "Structure and the Existence of the first sharp diffraction peak in amorphous germanium prepared in UHV and measured in-situ"

by: Petr Viščor

151/87 "DYNAMISK PROGRAMMERING"

Matematikprojekt af:
Birgit Andresen, Keld Nielsen og Jimmy Staal

Vejleder: Mogens Niss

152/87 "PSEUDO-DIFFERENTIAL PROJECTIONS AND THE TOPOLOGY OF CERTAIN SPACES OF ELLIPTIC BOUNDARY VALUE PROBLEMS"

by: Bernhelm Booss-Bavnbek
Krzysztof P. Wojciechowski

153/88 "HALVLEDERTEKNOLOGIENS UDVIKLING MELLEM MILITÆRE OG CIVILE KRÆFTER"

Et eksempel på humanistisk teknologihistorie
Historiespeciale

Af: Hans Hedal

Vejleder: Ib Thiersen

154/88 "MASTER EQUATION APPROACH TO VISCOUS LIQUIDS AND THE GLASS TRANSITION"

By: Jeppe Dyre

155/88 "A NOTE ON THE ACTION OF THE POISSON SOLUTION OPERATOR TO THE DIRICHLET PROBLEM FOR A FORMALLY SELFADJOINT DIFFERENTIAL OPERATOR"

by: Michael Pedersen

156/88 "THE RANDOM FREE ENERGY BARRIER MODEL FOR AC CONDUCTION IN DISORDERED SOLIDS"

by: Jeppe C. Dyre

157/88 " STABILIZATION OF PARTIAL DIFFERENTIAL EQUATIONS BY FINITE DIMENSIONAL BOUNDARY FEEDBACK CONTROL: A pseudo-differential approach."

by: Michael Pedersen

158/88 "UNIFIED FORMALISM FOR EXCESS CURRENT NOISE IN RANDOM WALK MODELS"

by: Jeppe Dyre

159/88 "STUDIES IN SOLAR ENERGY"

by: Bent Sørensen

160/88 "LOOP GROUPS AND INSTANTONS IN DIMENSION TWO"

by: Jens Gravesen

161/88 "PSEUDO-DIFFERENTIAL PERTURBATIONS AND STABILIZATION OF DISTRIBUTED PARAMETER SYSTEMS:

Dirichlet feedback control problems"

by: Michael Pedersen

162/88 "PIGER & FYSIK - OG MEGET MERE"

AF: Karin Beyer, Sussanne Blegaa, Birthe Olsen, Jette Reich , Mette Vedelsby

163/88 "EN MATEMATISK MODEL TIL BESTEMMELSE AF PERMEABILITETEN FOR BLOD-NETHINDE-BARRIEREN"

Af: Finn Langberg, Michael Jarden, Lars Frellesen

Vejleder: Jesper Larsen